2. Groups, Rings, and Fields

§2.1. Defining Addition and Multiplication

After a week of learning what numbers are, we can get into big kid math! We can finally talk about adding and multiplying two numbers! We'll have to wait for subtraction and division until later.

If we have two sets A and B, we can define the **disjoint union**, $A \sqcup B$, as the union of two sets A and B, where we "don't care" about duplicates. For example, if $A = \{1, 2, 3\}$ and $B = \{1, 2\}$, we would have that $A \cup B = \{1, 2, 3\}$, while $A \sqcup B = \{1, 1, 2, 2, 3\}$. Another way to think about this is to identify each element in $A \sqcup B$ from where it originally came from, so the element 1 from A and the element 1 from B are distinct in $A \sqcup B$ since they came from different places. You can think about $\{1, 1, 2, 2, 3\}$ as $\{1_A, 1_B, 2_A, 2_B, 3_A\}$

With this, we can define addition:

Definition 2.1.1

If we have two sets A and B, we can define the sum of their cardinalities as the cardinality of their disjoint union:

$$\#(A) + \#(B) = \#(A \sqcup B)$$

If we have two sets A and B, we can define their **Cartesian product**, $A \times B$, as the set of "(x,y) coordinates", where x is in the set A, and y is in the set B. For example, if $A = \{1, 2, 3\}$ and $B = \{1, 2\}$, we would have $A \times B = \{(1, 1), (1, 2), (2, 1), (2, 2), (3, 1), (3, 2)\}$.

With this we can define multiplication:

Definition 2.1.2

With two sets A and B, we can define the product of their cardinalities as the cardinality of their Cartesian product:

$$\#(A) \cdot \#(B) = \#(A \times B)$$

When we define new things like this, one thing we should check is if it's well defined with notions of equality that we've defined earlier. In this case, we should check if addition and multiplication are well defined with respect to bijections of sets. Check out the well-definedness section of the first chapter for a review of how to show addition is well defined with respect to bijections, and when you're ready, try doing the same for products on your own:

Exercise 2.1.3

Show that the product operation is well defined under bijections of sets.

This means that we can now talk about adding and multiplying numbers, instead of adding and multiplying the cardinalities of explicit sets.

Similarly, we can show a bunch of different properties of addition and multiplication:

- Commutativity/Associativity of addition
- Commutativity/Associativity of multiplication
- Distributivity

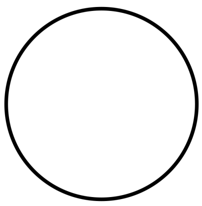
All of these come down to showing that there exists a bijection between two sets. For example, to show that the distributive law ($a \cdot (b + c) = a \cdot b + a \cdot c$) is true, we need to find a bijection between $A \times (B \sqcup C)$ and $(A \times B) \sqcup (A \times C)$, where $\#(A) = a, \#(B) = b, \#(C) = c$.

Exercise 2.1.4

Find the bijection and prove the distributive law

§2.2. Groups

Before we start directly talking about groups, let's talk a little bit about what abstract algebra is about. Abstract algebra is all about taking the concepts that we know about our "normal number system," the things that we learn in a middle/high school algebra class, and seeing what we can find by generalizing these ideas. By exploring what happens when we apply certain structures and restrictions on sets, we can find some ridiculously powerful tools to solve really interesting problems (check out the unconstructability of a 20 degree angle and the unsolvability of the quintic). We start with the notion of a group, and the best way to think of a group is that it describes the symmetries of something. Let's start off with the most symmetric thing we can think of, a circle, and try to describe the symmetries of it.

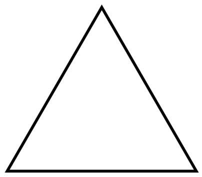


Imagine trying to move and flip this circle around so that it looks like we didn't do anything to it. We want our action on the circle to "preserve it's symmetries". There are a couple of things that we can do to this circle: we can rotate the circle about it's center by any amount, we can flip the circle about any diameter, and we can also do nothing. There are a couple of important things to notice:

- If we do one symmetry preserving action, and then do another symmetry preserving action, we've created some "composition of actions" that is also symmetry preserving actions. In other words, compositions of symmetry preserving actions is symmetry preserving.
- Composing actions is associative, as in if I do three actions a, then b, then c, it's the same as if I found the composed action of doing c after b, let's call it d, and did d after a. It's kind of interesting that the composition of actions isn't commutative (try flipping the circle about it's "x axis" and then rotating 90 degrees clockwise, and then doing this in reverse order).

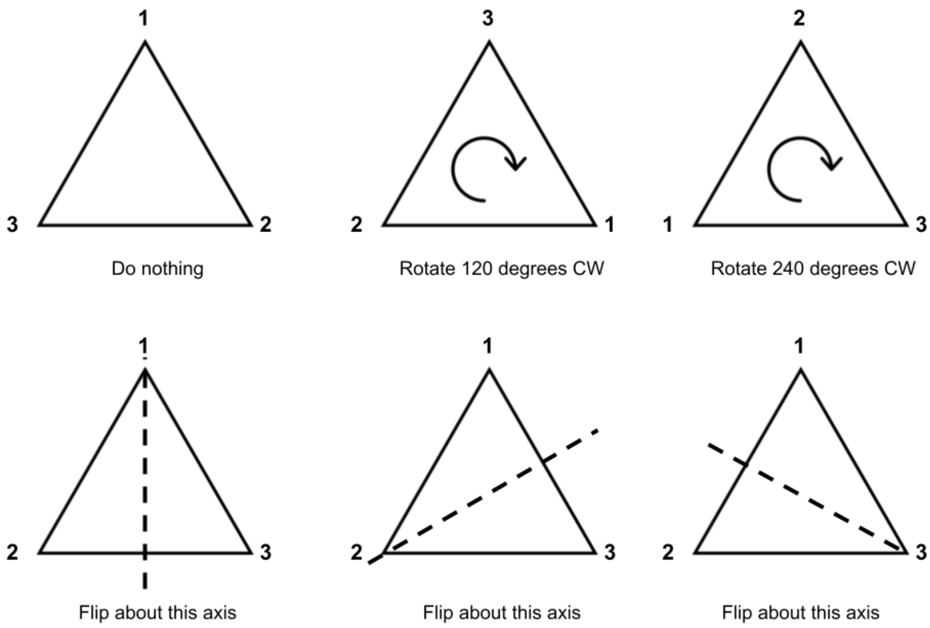
- Doing nothing is a symmetry preserving action.
- Any action we do has an action that we can do to "invert" what we did before. If we rotate our circle by 120 degrees, there exists the symmetry preserving action of rotating -120 degrees to undo our previous action.

Lets try messing around with some other shapes, how about a triangle:



A triangle doesn't have nearly as many symmetries as a circle. There are only a couple of things we can do to our triangle to make it look like where it started after our action:

- Doing nothing
- Rotating by 120 or 240 degrees
- Flipping our triangle about a median of the triangle



Notice we can't have more symmetries of the triangle, since any symmetry of a triangle permutes the vertices of a triangle uniquely, and there are a total of $3! = 6$ permutations of 3 vertices, so there can be a maximum of 6 symmetries of the triangle.

Even though a triangle's symmetries behave very differently from a circle, the properties we observed from the circle still hold:

- The composition of any two symmetry preserving actions is a symmetry preserving action. See if you can figure out what action we get if we first rotate 120 degrees clockwise, then do the leftmost flip in the picture above.

- Just as before, the composition action is associative.
- We have a "do nothing" action that preserves the symmetries of the triangle.
- Every action has an undo action. The undo action of a rotation is the opposite rotation, and the undo action of any flip is the same flip again.

We would like to create a "structure" on a set that helps capture the ideas that we've seen from playing with symmetries. We want elements of a set to represent symmetry preserving actions, and we want some way of composing these elements together to reach other elements in the set. Keeping this in mind, we define a group as follows:

Definition 2.2.1

A **group** is a set G with a binary operation $\star$ (which is just a function that takes in two elements from the set G and spits out another element of G) that follows these rules:

- The set is **closed** under $\star$, meaning that for any two elements $a, b \in G$, we have that $a \star b \in G$ (which is garunteed since $\star$ is a function from $G \times G$ to G)
- The $\star$ operation is **associative**, as in $a \star (b \star c) = (a \star b) \star c$
- The set has an **identity element**, an element $e \in G$ such that for any $a \in G, e \star a = a \star e = a$ (this is the "do nothing" operation we saw earlier)
- Every element $a \in G$ has an **inverse** element a^{-1} such that $a \star a^{-1} = a^{-1} \star a = e$ (this is the idea that every action has a corresponding "undo action")

Hopefully this section helped motivate why we define a group as we do.

As practice for working with groups, we'll prove a property of groups here and leave one for you to try as an exercise:

Example 2.2.2

Show that for a group G, if an element g has a right inverse r (as in $r \star g = e$) and has a left inverse l (as in $g \star l = e$), then $r = l$.

Proof. We start with $r \star g = e$. "Multiplying" by l on the left on both sides (note that we need to be careful about what side we multiply on, since "multiplication" isn't necessarily commutative) gives us

$$r \star g \star l = e \star l$$

Applying properties of identity and associativity gives us

$$r \star (g \star l) = l$$

And using the fact that $g \star l = e$ gives us $r = l$, as desired. □

Exercise 2.2.3

Show that the identity element in a group G is unique, i.e. if $g \star e = e \star g = g$ and $g \star e' = e' \star g = e'$ for all $g \in G$, then $e = e'$

§2.3. Examples of Groups

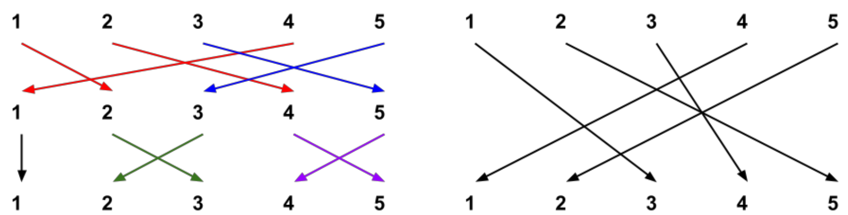
Gunning does not give nearly enough examples of groups, so here's a bunch to keep in mind.

- **The Cyclic Group**: The set $\{0, 1, 2, ..., n-1\}$ forms a group of order n (where the order of a group is just the cardinality of it's set), with the binary operation "addition modulo n". You can verify that this satisfies the group axioms. Note that addition modulo n is a commutative operation (as in $a + b \pmod{n} = b + a \pmod{n}$), so this is an **abelian group** (or commutative group).

  You can also think about this group as the set of rotations that preserve the symmetries of a regular n-gon: $k \cdot 360/n$ degree rotations for $k \in \{0, 1, ..., n-1\}$, with the binary operation "composition of rotations". These groups represent the same thing and act the same way, so we call these two groups **isomorphic** (we'll formalize this in a bit).

- **The Dihedral Group**: The group D_n is the set of rotations and flips that preserve the symmetries of a regular n-gon, with the binary operation "composition of actions". The triangle group we defined last section is usually denoted as D_3. You can work through the details yourself, but D_n has $2n$ elements (1 "do nothing" action, $n-1$ rotations, and n reflections). This group is not abelian, and to convince yourself of this, go back to the picture of triangle symmetries and check to see that a rotation and a flip is not the same as the flip first and then the rotation.

- **The Symmetry Group**: The group S_n is the set of all permutations of n elements. One nice way to visualize this action is to imagine a regular n-gon without it's edges, only the vertices. You can map this set of vertices back to itself just by shuffling around the vertices any way you want. Another way to visualize this action is shown below:

  We can visualize permutations of the set $\{1, 2, 3, 4, 5\}$ by drawing arrows from one number to wherever the permutation maps it. The composition of two permutations on the left is equal to the single permutation on the right.

  Note: S_3 and C_6 are the only groups of order 6. If you're wondering why D_3 isn't on the list, it's isomorphic to S_3. Go try to convince yourself why that is by looking at the triangle symmetries in the previous section.

- **The Klein 4 Group**: The group K_4 is the set $0, a, b, a \star b$, where $\star$ acts like this:
  - 0 is the identity element wrt $\star$
  - $a \star a = b \star b = (a \star b) \star (a \star b) = 0$
  - $a \star b = b \star a$

There are many nice ways to think about this group. One way is to see it as the symmetries (rotations and reflections) of a rectangle, and another way to see it is as the group $C_2 \times C_2$ (this is a product group, where each coordinate acts like an element in a cyclic group of order 2). Gunning writes out the Cayley table for this.

Note: K_4 and C_4 are the only groups of order 4. It's kind of cool that all groups of order 4 are abelian.

- **The Trivial Group** This group only has the identity element e. It trivially satisfies the group axioms.

Although totally unrelated, here are some fun groups to check out if you're curious (or if you're really interested, take MAT345):

- The Quaternion Group (an order 8 group)
- The orientation-preserving symmetries of a tetrahedron (an order 12 group)
- The orientation-preserving symmetries of a hypercube (an order 48 group)
- The set of all automorphisms of a group/field (we'll define automorphisms later)
- The Fundamental Group (some neat topology stuff, check out Munkres or MAT365 for more)

§2.4. Homomorphisms and Isomorphisms

Isomorphisms are more easily motivated so we'll discuss them first. Just as we had a notion of "sameness" of sets through bijections, we would like a way to talk about when two groups are the same. However, an unrestricted bijection isn't enough now to say two groups are the same, since they don't respect the structure of groups. For instance, although C_4 and K_4 have the same order, they clearly act differently. C_4 has an element of order 4 (when we talk about an element a having order n we mean that n is the smallest number such that $\underbrace{a \star a \star \ldots \star a}_{n \text{ times}}$ equals the identity element, while K_4 only has elements of order 1 and 2. So clearly, the internal structure of these groups are different.

Definition 2.4.1

An **group isomorphism** φ from a group G (with operation $\star_G$) to a group H (with operation $\star_H$) is a **bijective** function between the sets G and H that must respect the group operations in the following way:

- $\varphi(g_1 \star_G g_2) = \varphi(g_1) \star_H \varphi(g_2)$ for $g_1, g_2 \in G$. What this means is that composing two elements in G and then mapping to H gives you the exact same result as if you map both elements to H and then compose the two elements. Notice that the operations $\star_G$ and $\star_H$ corresponding to groups G and H might be different.
- $\varphi(e_G) = e_H$, which just means that φ maps the identity of G to the identity of H. Note: this last bullet is actually implied by the first bullet, but it helps to think about this as a property of isomorphism

I think that this commutative diagram helps to understand what's going on (a commutative diagram is just a diagram where paths with the same start and end points lead to the same result). Let $g_1 \star_G g_2 = g_3$ and $h_1 \star_H h_2 = h_3$. Then we have:

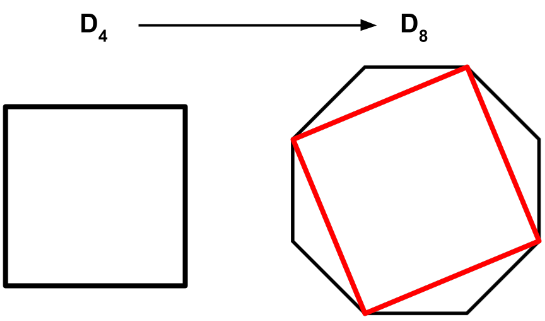
$$\begin{tikzcd}
(g_1, g_2) \arrow[r, "\varphi"] \arrow[d, "\star_G"] & (h_1, h_2) \arrow[d, "\star_H"] \\
g_3 \arrow[r, "\varphi"] & h_3
\end{tikzcd}$$

So since isomorphisms force an element g and it's image $\varphi(g)$ to act in the same way, an isomorphism can only exist between two groups if they act the same way everywhere.

Definition 2.4.2

A **group homomorphism** between two groups G and H is a function $\phi : G \to H$ following the same properties that an isomorphism does (so the function must preserve the group operations of G and H), with the relaxation that the function does not necessarily need to be bijective. A bijective homomorphism is precisely an isomorphism.

A nice picture of homomorphism to keep in mind is a homorphism ψ from D_4 (the symmetries of a square) to D_8 (the symmetries of an octagon), by mapping the symmetries of a square to symmetries of an octagon that preserve symmetries of an inscribed square.



§2.5. Rings, Integral Domains, and Fields

Rings and Fields are significantly less intuitive, and there are a lot less examples that are immediately approachable, so our examples of rings and fields will be somewhat generic.

Definition 2.5.1

The idea behind a ring is that instead of having one binary operation on a set, we have two, denoted by $+$ and $\times$. A ring R must satisfy the following rules:

- R is an abelian group under $+$ (so it's closed under $+$, $+$ is commutative and associative, $+$ has an identity, and all elements of R have inverses with respect to $+$).
- R has an identity element with respect to $\times$ (the equivalent of the number 1 in $\mathbb{R}$), R is closed under $\times$, and $\times$ is associative and commutative. Note that the only thing stopping R from being an abelian group under $\times$ is that it doesn't necessarily have inverses with respect to $\times$.
- $+$ and $\times$ follow the distributive property, so for $a, b, c \in R$, we have $a \times (b+c) = a \times b + a \times c$

We call the $+$ operation addition and the $\times$ operation multiplication, to match our intuition of what these symbols mean and do. We denote the additive inverse of an element a as $-a$, and the multiplicative inverse of an element a (if it has one) as a^{-1}. We also sometimes just write the product $a \times b$ as ab.

Multiplication by 0 in a ring acts just like you think it should, since by the distributive property, $a = a \times 1 = a(1+0) = a + a \times 0$, so adding $-a$ on both sides (where it doesn't matter where we add $-a$ now since addition is commutative) we get that $a \times 0 = 0$.

Try using a similar argument to show this:

Exercise 2.5.2

Show that if $a, b \in R$, $(-a)(b) = -(ab)$

Removing the restriction for multiplicative inverses (inverses wrt $\times$) leads to some rings acting very strangely. For example, lets look at the set $0, 1, 2, 3, 4, 5$, now viewed as a ring, with addition defined as addition mod 6, and multiplication defined as multiplication mod 6. This satisfies the definition of a ring (since the modulo operation is well defined over addition and multiplication, and addition and multiplication sastisfy the ring axioms over the ring of integers (Z)). However, strange things start to happen that we don't usually see in things like $\mathbb{Z}$, $\mathbb{Q}$, or $\mathbb{R}$: multiplying two nonzero elements sometimes gives us zero. For example, $2 \times 3 \pmod 6 = 0$. We call pairs of numbers like this **zerodivisors**. Zerodivisors cannot have multiplicative inverses, since if they did, $2 \times 3 \times 2^{-1} = 0 \times 2^{-1}$ implies $3 = 0$.

If we add the extra restriction on a ring that there cannot not be any zerodivisors, we get what we call an integral domain.

Definition 2.5.3

An **integral domain** is a special case of a ring, where for elements a, b in our ring, we have that $a \times b = 0$ implies $a = 0$ or $b = 0$.

A common example of an integral domain is the integers, $\mathbb{Z}$. You can see that there are no zerodivisors in $\mathbb{Z}$, but we still see that there are numbers in $\mathbb{Z}$ that don't have multiplicative inverses: for instance 2 has no multiplicative inverse in $\mathbb{Z}$ since $\frac{1}{2}$ is not an integer.

Finally, if we add on the restriction that any nonzero element must have a multiplicative inverse (zero can't have a multiplicative inverse since anything times zero is zero), we get what we call a field.

Definition 2.5.4

A bf is a set F equipped with addition $(+)$ and multiplication $(\times)$ such that:

- F is an abelian group over $+$.
- $F \sim 0$ is an abelian group over $\times$.
- $+$ and $\times$ follow the distributive property.

$\mathbb{Q}$, $\mathbb{R}$, and $\mathbb{C}$ are all great examples of fields to keep in mind.

Homomorphism and isomorphisms of rings and integral domains are slightly different than group homomorphims, since while they also have to preserve the group structure given by $+$, it must also preserve the multiplicative identity. So we require for a a homomorphism φ between rings R and S:

- $\varphi(r_1 +_R r_2) = \varphi(r_1) +_S \varphi(r_2)$, the requirement preserving the group structure with respect to addition
- $\varphi(0_R) = 0_S$, the requirement preserving the additive identity
- $\varphi(1_R) = 1_S$, the requirement preserving the multiplicative identity

And an isomorphism of rings is just a bijective homomorphism. A homomorphism or isomorphism of fields must also preserve the group structure of $F \sim 0$ with respect to $\times$, so we require the additional constraint that $\varphi(r_1 \times_R r_2) = \varphi(r_1) \times_S \varphi(r_2)$.

Gunning talks about polynomial rings for a bit: $R[x]$ is a ring of polynomials with coefficients in the ring R, where adding and multiplying polynomials acts exactly like you think it would (add coefficients of equal powers of x, multiply by foiling). As a final, more challenging exercise for this section, consider the ring $F[x]$, where F is a field.

Exercise 2.5.5

Show that $F[x]$ is not a field, but that $F[x]$ is an integral domain.

§2.6. Constructing the Integers and the Rationals

Gunning walks through construction of the Integers and Rationals, but doesn't provide much motivation for why we define the equivalence classes the way he does. We'll provide some motivation for this, but not go through all the details Gunning does.

In defining the integers, we need to somehow define the "negative numbers" from working with the natural numbers. To do this, we would like to consider the difference "a-b" (which we haven't actually defined yet) as an element (a, b) in $\mathbb{N} \times \mathbb{N}$. We would like the integers to form a ring, so we need to define the $+$ and $\times$ operations on it. Secretly, we think of adding (a, b) and (c, d) as adding $(a - b)$ and $(c - d)$. Since we want $(a - b) + (c - d) = (a + c) - (b + d)$, we define addition on $\mathbb{N} \times \mathbb{N}$ as

$$(a, b) + (c, d) = (a + c, b + d)$$

Similarly, for multiplication, we secretly think of multiplying (a, b) and (c, d) as adding $(a - b)$ and $(c - d)$. Since we want $(a - b)(c - d) = (ac + bd) - (ad + bc)$, we define multiplication on $\mathbb{N} \times \mathbb{N}$ as

$$(a, b) \times (c, d) = (ac + bd, ad + bc)$$

To turn these coordinates (a, b) into numbers $a - b$, we need to "quotient out the ambiguity", since right now coordinates like $(2, 4)$ and $(3, 5)$ define the same number. We want to define an equivalence relation such that $(a, b) \asymp (c, d)$ when $a - b = c - d$, which only happens when $a + d = b + c$. Note that although this equivalence relation represents the negative numbers, we've defined them only using the natural numbers. That's kind of neat! By quotienting out by this equivalence relation, we get $\mathbb{N} \times \mathbb{N}/\asymp$, which is a ring and acts just like we would like the integers to. Gunning works through the rest of the details, and leaves it up to you in the homework to show that this equivalence relation is well defined with respect to the addition and multiplication we defined.

To extend the integers (an integral domain) to the rationals (a field), we need to fulfill the last axiom of being a field: we need to adjoin multiplicative inverses to all non-invertible elements of the integers. Just as we secretly used subtraction to motivate our definition of the integers, we'll secretly use division to motivate our definition of the rationals. We want represent the fraction $\frac{a}{b}$ by a coordinate $(a, b) \in \mathbb{Z} \times (\mathbb{Z} \sim 0)$, where we remove the possibility of b being zero since we don't want zero to have an inverse (this would conflict with the concept that anything times zero is zero). Since adding two fractions $\frac{a}{b}$ and $\frac{c}{d}$ gives $\frac{ad+bc}{bd}$, we define addition on $\mathbb{Z} \times (\mathbb{Z} \sim 0)$ as

$$(a, b) + (c, d) = (ad + bc, bd)$$

Since multiplying two fractions $\frac{a}{b}$ and $\frac{c}{d}$ gives $\frac{ac}{bd}$, we define multiplication on $\mathbb{Z} \times (\mathbb{Z} \sim 0)$ as

$$(a, b) \times (c, d) = (ac, bd)$$

And to turn these coordinates (a, b) into fractions $\frac{a}{b}$, we need to "quotient out the ambiguity" of fractions like $\frac{3}{4}$ and $\frac{6}{8}$ and $\frac{-3}{-4}$. We want to define an equivalence relation such that $(a, b) \asymp (c, d)$ when $\frac{a}{b} = \frac{c}{d}$, which only happens when $ad = bc$ By quotienting out by this equivalence relation, we get $\mathbb{Z} \times (\mathbb{Z} \sim 0)/\asymp$. This is a field, since all nonzero equivalence classes (a, b) have a multiplicative inverse (b, a) (since (ab, ab) is in the same equivalence class as $\frac{1}{1}$, the multiplicative identity of $\mathbb{Z} \times (\mathbb{Z} \sim 0)$), and it acts just like we would like the rationals to act.

What we've just constructed is called a "fraction field" of the integers, and this method of adjoining inverse elements works to turn any integral domain into a field. We can't do the same process for rings with zerodivisors, since if fractions like $\frac{1}{2}$ and $\frac{1}{3}$ were multiplied, we would get a fraction with zero in the denominator.

§2.7. Ordered Rings

The integers and rationals have an order: 2 is less than 3, and $\frac{8}{5}$ is greater than $\frac{-1}{4}$. We would like to generalize this concept and determine when we can put an order on a ring.

The statement $a < b$ for integers a and b can be rewritten as $a - b < 0$, and the statement $a > b$ can be rewritten as $a - b > 0$. Since 0, the additive identity, is contained in all rings, we should define an order on a ring in terms of 0 rather than the elements of a particular ring. To say something is greater than or less than zero in the integers, we have a notion of positive and negative numbers. If $a - b$ is in the set of positive integers, we know $a > b$. Notice that positive and negative numbers are "opposites" with respect to additive inverses: if a is positive, $-a$ is negative, and if $-a$ is positive, a is negative. Also notice that the positive integers are closed under addition and multiplication (which is not implied by anything we've said so far). With these ideas in mind, we define an ordered ring as follows:

Definition 2.7.1

An **ordered ring** is a ring R with a subset $P \subset R$ such that:

- For any $a \in R$, exactly one of $a = 0$, $a \in P$, or $-a \in P$ is true.
- P is closed under $+$ and $\times$

Although it may seem like these conditions aren't sufficient for an ordered ring to act as closely to something like the integers and rationals as we'd like it to, these conditions are enough for us to define a partial order on the set R, where $a \geq b$ if and only if $a - b \in P$ or $a - b = 0$. You can check to see that these conditions alone are enough to satisfy the properties of transitivity, antisymmetry, and reflexivity. This is a special type of order, called a total order, where any two elements in the ring can be compared. Not all partial orders are total orders; for an example, consider the partial order on the power set of some set A, where for subsets B and C of A, $B \geq C$ if and only if C is a subset of B. If $B \cap C \neq B$ or C, then we can't say anything about the order of B and C.

§2.8. The Real Numbers and The Archimedian Property

In this chapter, we've formed the integers (an ordered integral domain) from the natural numbers (which don't even form a group), and we've formed the rationals (an ordered field) from the integers. The real numbers also form an ordered field, so what property distinguishes the rationals from the reals? This property is called **completeness**, and in Chapter 2 we'll show how we can construct the real numbers by adding this property to the rationals.

Definition 2.8.1

On a set S with a total order (where any two objects x and y in S are comparable), S is defined as **complete** if any subset $U \subset S$ with an upper bound b has a **least upper bound** b' such that:

- b' is an upper bound for U
- For any upper bound b of U, $b' \leq b$

We denote b' as the supremum of U, written as $\sup(U)$. An ordered field with this completeness property is unsuprisingly called a **complete ordered field**, and we leave it as an exercise to show that a complete ordered field garuntees the existance of a greatest lower bound on a subset U (assuming U is bounded below), denoted by $\inf(U)$ (the infimum of U).

Exercise 2.8.2

Show that a complete ordered field has the greatest lower bound property by finding a way to relate supremums to infimums.

For a quick example, $\mathbb{Q}$ is not a complete ordered field. For example, take the infinite subset $\{3, 3.1, 3.14, 3.141, ...\}$ of $\mathbb{Q}$that approaches π from below. But for any rational upper bound we can find for this set , there exists a slightly smaller rational that is still an upper bound for this set (you can take this on intuition, or use the upcoming theorem).

Gunning shows how the rationals can be viewed as a subset of the reals by showing an injection φ from $\mathbb{Q}$ to $\mathbb{R}$, which allows us to compare rational and real numbers. This leads to an extremely important theorem that we would like to highlight:

Theorem 2.8.3

Gunning Theorem 1.6 Part 1: For any real number $a > 0$ there is are integers n_1 and n_2 with the property that $n_1 > a$ and $\frac{1}{n_2} < a$

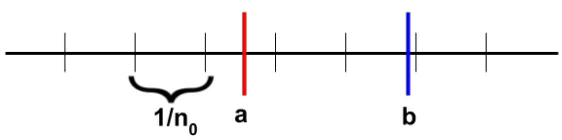
Gunning Theorem 1.6 Part 2: For any real numbers $a < b$ there is a rational number r with the property that $a < r < b$

These two properties together form the **Archimedian property** of the real numbers.

Proof. Here's an outline of how to think through this proof:
Part 1:

- To show there exists an $n_1 > a$ directly is hard, since we can't really do much with the upper bound properties of the reals (since assuming the set $\{a\}$ has an upper bound might assume the theorem already). Instead, we try showing a contradiction.
- If we assume a is larger than all integers, then a is an upper bound for $\mathbb{Z}$. Let a' be the least upper bound for $\mathbb{Z}$ We now want to use some property of integers to contradict this least upper bound. What we can use is the inductive property of the integers: if x is in $\mathbb{Z}$, then so is $x + 1$.
- Since a' is the least upper bound of $\mathbb{Z}$, there exists some integer x such that $x > a' - 1$. Then, $x + 1$ is an integer but $x + 1 > a'$ which contradicts the fact that a' is a least upper bound.
- Notice that in this proof we used both assumptions that $\mathbb{Z}$ has an upper bound and a least upper bound, making full use of completeness.
- we know that reciprocals flip inequality signs, so finding an n_2 greater than $\frac{1}{a}$ immediately gives us an n_2 such that $0 < \frac{1}{n_2} < a$.

Part 2: the best way to see this is with a picture:



From part 1, we know there exists an integer n_0 such that $\frac{1}{n_0} < b - a$. Then we can hop up the number line in $\frac{1}{n_0}$ increments until we fall in the interval (a, b) (which we're garunteed to since our jump sizes are smaller than the distance between a and b).

Notice that this part of the proof tells us two things: that there is a real number between any two rational numbers but also that there is a rational number between any two real numbers. This is an extremely important fact to keep in mind as you work towards Chapters 2 and 3. $\square$

§2.9. Key Takeaways

The most important thing to walk away from this chapter from is an understanding of the thought processes we worked through to develop new tools from old ones, like how we constructed the integers and rationals from the natural numbers. Keep a couple examples of groups, rings, integral domains, and fields with you to gain an intuition of what these objects "feel like." And finally, keep completeness and the archimedian property in the back of your head, they will show up A LOT later.

3. Vector Spaces

Linear algebra speedrun time woooooooo! The 2022 216 section did this in 3 days, let us know if you beat our record :)

§3.1. From Fields to Vector Spaces

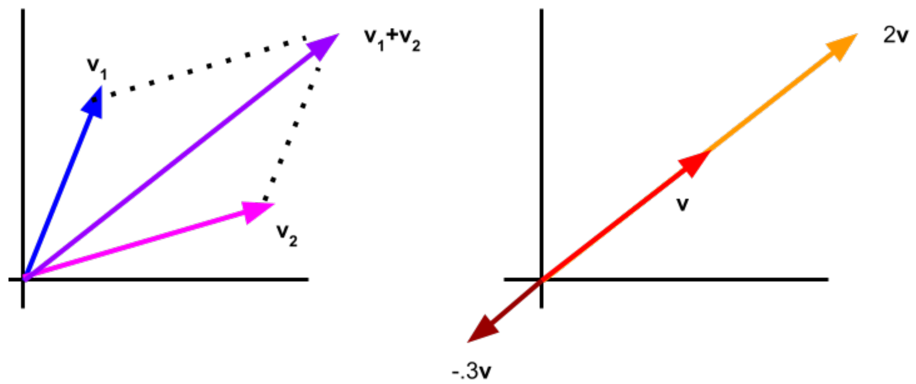
Vector spaces are an algebraic structure, just like groups, rings, and fields. We'll denote vectors in this vector space with boldface font, to distinguish them from scalars.

Definition 3.1.1

A **vector space** is an abelian group V (the vectors) over a field F (the scalars). In this vector field, we have addition of vectors (given by the addition defined on the abelian group V), as well as a scalar multiplication between elements of F and V that must satisfy the following properties:

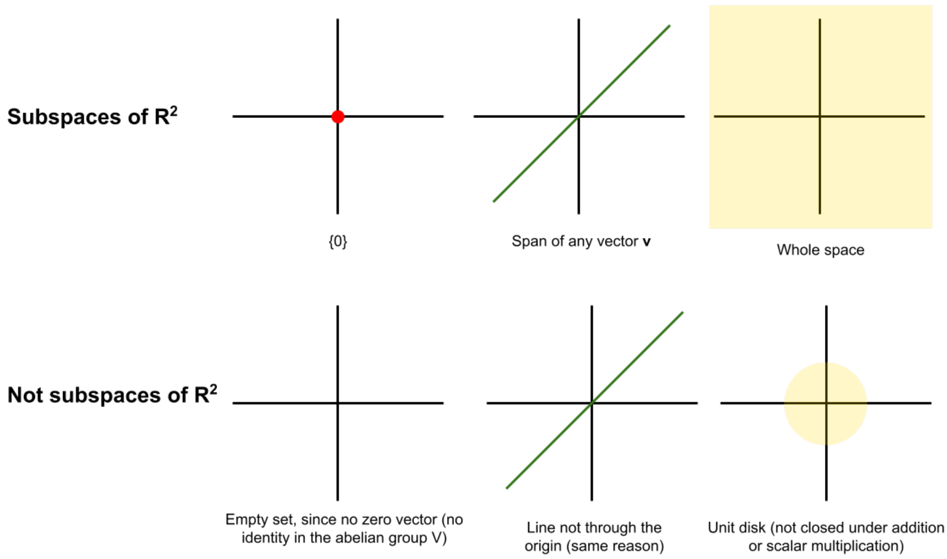
- the scalar multiplication is associative: $a(b\mathbf{v}) = (ab)\mathbf{v}$ for $a, b \in F, \mathbf{v} \in V$
- the multiplicative identity of F (which we denote as 1) is also the identity for scalar multiplication: $1\mathbf{v} = \mathbf{v}$ for all $\mathbf{v} \in V$
- the distributive law holds: $(a+b)(\mathbf{v}) = a\mathbf{v} + b\mathbf{v}$ and $a(\mathbf{v}_1 + \mathbf{v}_2) = a\mathbf{v}_1 + a\mathbf{v}_2$

The standard example of a vector space is F^n (over the field F), who's vectors are the set of n-tuples of elements in F. To get a better understanding of what addition and scalar multiplication are in F^n, lets look at $\mathbb{R}^2$. Addition, numerically, involves adding two 2-tuples of real numbers, (x_1, y_1) and (x_2, y_2) coordinate wise. Pictorially, this looks like drawing vectors from the origin to each point, and then moving the tail of one to the tip of the other. Multiplication of a scalar a and a vector (x_1, y_1) numerically is just multiplying each coordinate of (x_1, y_1) by a. Pictorially, this is just scaling the vector from the origin to (x_1, y_1) by a factor of a. Putting these rules on F^n defines a vector space just how you would expect it.



§3.2. Subspaces

Subspaces are defined exactly as you would think. Gunning provides the definition, and we provide a couple of examples:



Another good example of a vector space to keep in mind is the polynomial ring $\mathbb{R}[x]$ (which is an abelian group) over the field $\mathbb{R}$.

- The set of all polynomials in $\mathbb{R}[x]$ of degree 2 or less is a subspace, since it contains the "zero vector" (the polynomial 0), and is closed under addition and scalar multiplication.
- The set of all polynomials $p \in \mathbb{R}[x]$ such that $p(5) = 0$ is a subspace, since it contains 0 and is closed under addition and scalar multiplication.
- The set of all polynomials $p \in \mathbb{R}[x]$ such that $p(0) = 5$ is a subspace, since it doesn't contain 0 (and it's also not closed under addition and scalar multiplication).

§3.3. Quotient Spaces

Given a vector space V and a subspace V_0, we can define an equivalence relation on vectors by setting $\mathbf{v}_1 \asymp \mathbf{v}_2 \pmod{V_0}$ whenever $\mathbf{v}_1 - \mathbf{v}_2 \in V_0$. By quotienting V by V_0, we "glue" all vectors in the same equivalence class together, and treat them as one element in V/V_0. In other words, the equivalence classes are the elements of the abelian group in the vector space V/V_0. To actually prove that a quotient space is a vector space, we need to show that addition and multiplication are well defined over these equivalence class, and that they follow the axioms of a vector space

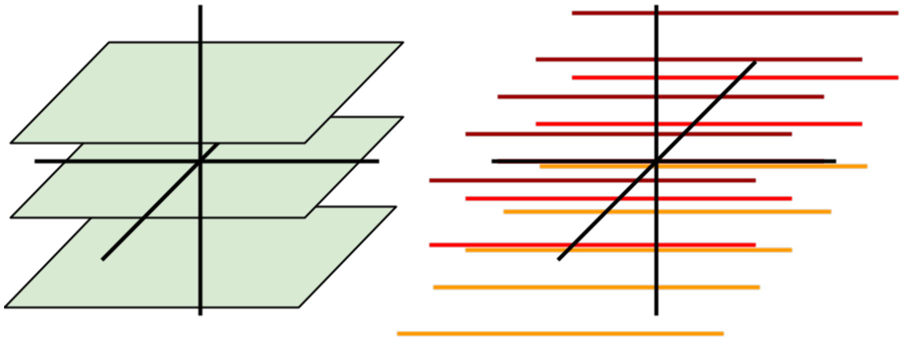
Exercise 3.3.1

Show that addition and scalar multiplication are well defined over the equivalence classes defined above, and use this to show that a quotient space has the algebraic structure of a vector space.

It helps a lot to think about what quotienting looks like visually, and what elements of V/V_0 look like. Lets look at the vector space $\mathbb{R}^3$, and the subspace $\mathbb{R}^2$ (the x-y plane). Two vectors in $\mathbb{R}^3$ are glued together if and only if their z components are the same, so equivalence classes of $\mathbb{R}^3$ (or elements of $\mathbb{R}^3/\mathbb{R}^2$) look like planes parallel to the x-y plane. The identity element of $\mathbb{R}^3/\mathbb{R}^2$ is the plane containing the identity in $\mathbb{R}^3$, which is just the x-y plane. This space "looks like" (is in bijection with) a line, and this should make sense since we're quotienting a 3-dimensional space by a 2-dimensional space, collapsing it into something that looks like a 1-dimensional space. We'll define dimension more rigorously in a bit.

If we quotient $\mathbb{R}_3$ by a line (lets say the line $y = 0$ in the x-y plane), then two vectors are equivalent to each other if and only if both their y and z coordinates are the same. The identity element of $\mathbb{R}^3$/y=0 is just the line $y = 0$. This space is in bijection with a plane, which makes sense since we're quotienting a 3-dimensional space with a 1-dimensional space, collapsing it into something that looks like a 2-dimensional space.

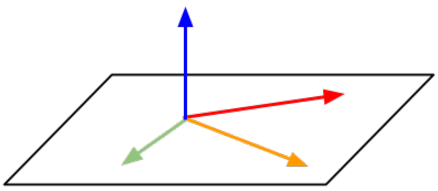
Both of our examples are shown in pictures below.



§3.4. Basis and Dimension

This section will cover some of the ideas related to span, linear independence, and linear dependence, and some of the theorems that go along with it.

Linear independence is the idea that "you can reach a vector in only one way." Given a span of vectors, span$(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$, linear independence says that the only way to make the zero vector by scaling and adding vectors in this span is to multiply every $\mathbf{v}_i$ by zero. More rigorously, this says that $\sum_{i=1}^{n} a_i \mathbf{v}_i = 0$ if and only if $a_i = 0$ for all $0 \leq i \leq n$. This is equivalent to the statement that any vector can be reached in exactly one way, since if a vector $\mathbf{v}$ can be reached in several ways, then subtracting by $\mathbf{v}$ gives us multiple ways to reach the zero vector. It's helpful to visualize linear independence by looking at examples in $\mathbb{R}^3$

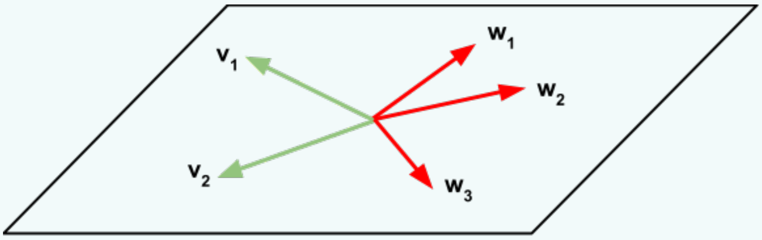


The red, orange, and green vectors are linearly independent, since you can add scaled combinations of red and orange to get green. However, all of the vectors are pairwise linearly independent, and the set of all four vectors are linearly independent. Although we haven't defined "dimension" yet, the intuition you should get is that two linearly independent vectors span a 2D space, three linearly independent vectors span a 3D space, and so on. You should also begin to notice the idea that a 2D space cannot have a set of 3 or more vectors that are linearly independent, and more generally, an n-dimensional space cannot have a set of $n+1$ or more vectors that are linearly independent.

Now we'll discuss the motivation and proofs for two important theorems in this chapter:

Theorem 3.4.1

The Basis Theorem (Gunning Theorem 1.7): If $\{\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m\}$ is in $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$, and $m > n$, then $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_m$ are linearly dependent.



Proof. This is just a formalization of our intuition from earlier. We just need to figure out how to rigorously argue the thought that "an n-dimensional space cannot have a set of $n+1$ or more vectors that are linearly independent." Our intuition tells us that we can fit n linearly independent vectors inside "an n-dimensional space", so if we can use find n $\mathbf{w}_i$ vectors that span the same space as the n $\mathbf{v}_i$ vectors, we have a great start. And that's actually enough to complete the proof, since if $\mathbf{w}_{n+1}$ is in $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$, it's also in $\text{span}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n)$, which shows that the set of $\mathbf{w}_i$ vectors is linearly independent. This idea of "building a basis" will show up a lot.

Let's assume all of the $\mathbf{w}_i$ vectors are linearly independent. This implies that none of the $\mathbf{w}_i$ are zero (convince yourself why that is). Let's start fitting the $\mathbf{w}_i$ into the span of the $\mathbf{v}_i$s. We know that $\mathbf{w}_1$ is in $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$, so we have that $\mathbf{w}_1 = a_1\mathbf{v}_1 + a_2\mathbf{v}_2 + \cdots + a_n\mathbf{v}_1$ for some scalars $a_1, a_2, \dots, a_n$. By isolating $\mathbf{v}_1$, we get that $\mathbf{v}_1 = b_1\mathbf{w}_1 + b_2\mathbf{v}_2 + \cdots + b_n\mathbf{v}_1$ for some scalars $b_1, b_2, \dots, b_n$, where $b_1 \neq 0$ (we can assume this without loss of generality, and if all the b_is were zero, $\mathbf{w}_1$ would be zero). We were given that $textbfw_1 \subset \text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$, and we showed that $\mathbf{v}_1 \subset \text{span}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n)$, so we get that

$$\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = \text{span}(\mathbf{w}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$$

Intuitively, what we've done is we found some $textbfv_1$ that had a component in the direction of $textbfw_1$, so by replacing $textbfv_1$ with $textbfw_1$ we are not losing any space. Since $textbfw_1$ is in the span of all of the $textbfv_i$s, we are not gaining any space either, so the space stays the same.

We can repeat this step of swapping $textbfv_1$ with $textbfw_1$ to swap $textbfv_2$ with $textbfw_2$, and keep going until we show that $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n) = \text{span}(\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n)$. As we discussed earlier, this completes the proof since we can write $\mathbf{w}_{n+1}$ as a linear combination of $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_n$, giving us a contradiction. □

A lot of linear algebra is visual, so thinking about concepts informally in terms of dimensionality and "fitting vectors in spaces" is often very helpful.

Gunning goes on to define a basis (a linear independent spanning set). Here's a quick exercise to make sure you understand the idea of bases and the purpose of the basis theorem:

Exercise 3.4.2

Using the basis theorem and some reasoning about linear independence, reason why the number of vectors in a basis is unique.

Based off our intuition of the number of vectors in a space and our concept of dimensionality of the space, it makes sense why the dimension of a vector space is defined as the number of vectors in a basis for this space.

Next, we will formalize another one of our ideas from earlier:

Theorem 3.4.3

Gunning Theorem 1.8: If V is a (finite dimensional) vector space and W is a subspace of V, then

$$\dim V = \dim W + \dim V/W$$

Proof. Rearranging our goal, we get that $\dim V - dimW = dimV/W$, which is exactly what our intuition told us when we were working with quotient spaces. When we quotiented $\mathbb{R}^3$ by a one dimensional space, we got a two dimensional space, and when we quotiented it by a two dimensional space, we got a one dimensional space. This statement is just a formalization of our "dimension collapse" idea from earlier.

We want to look at the dimensions of vector spaces, so we should definitely pick bases for V and W. Since W is a subspace of V, we want to relate W as much as we can to V, so let's try to pick bases for V and W that are related. Let the dimension of V be m and the dimension of W be n (where $m \geq n$). We need somewhere to start, so let's choose a basis for W, call it $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_n$ (we start with the smaller space since it turns out to be easier than starting with the larger space, oftentimes you have to mess around with a couple of ideas before you come across one that works).

We now build V from W (this is a very common technique for relating bases between a space and a subspace of itself). If $W \neq V$, there is at least one vector $\mathbf{v}_1$ in V but not in W, so this vector is linearly independent to all of the $\mathbf{w}_i$s. If $\text{span}(W, \mathbf{v}_1) \neq V$, we can find some $\mathbf{v}_2 \in V$ that is not in $\text{span}(W, \mathbf{v}_1)$ and is therefore linearly independent to the $\mathbf{w}_i$s and $\mathbf{v}_1$. The basis theorem forces this to stop once we adjoin $\mathbf{v}_{m-n}$, so in this way, we've built a basis $\{\mathbf{w}_1, \ldots, \mathbf{w}_n, \mathbf{v}_1, \ldots, \mathbf{v}_{m-n}\}$ for V.

All that's left to do is to show that $\dim V/W = m - n$, and to do this we need to find a basis for V/W. Intuition carries us almost the rest of the way to the end of the proof. Looking back at the pictures in the quotient spaces section, we see that equivalence classes of V/W look like copies of W that span V by going in the directions that W doesn't. For example, looking at $\mathbb{R}^3$ quotient the line $x = 0$, we get lines parallel to $x = 0$ spanning the y and z directions.

Equivalence classes in V/W are of the form $\mathbf{v} + W$ for $\mathbf{v} \in V$. This is an extremely important idea to understand. To see this, consider what equivalence class $\mathbf{v}_1$ lies in.

Any vector $\mathbf{w} + \mathbf{v}_1$ for $\mathbf{w} \in W$ is clearly in this equivalence class, and that is all of the vectors in this equivalence class (see if you can reason this yourself using the transitivity of equivalence classes). Take a couple minutes to make sure you fully understand what is going on here, this is what makes quotient spaces so tricky.
Combining our understanding of equivalence classes, and that we want to make a basis V/W that pushes equivalence classes in directions that W doesn't cover, we note that $\mathbf{v}_1, \ldots, \mathbf{v}_{m-n}$ are linearly independent vectors that go in the directions we need W to go, and we guess a basis for V/W of the form $\{\mathbf{v}_1 + W, \ldots, \mathbf{v}_{m-n} + W\}$. Again, make sure you understand where this idea comes from. We leave it up to you to show that this is a basis for V/W, as this is a very useful exercise to solidify many of the concepts we've already discussed.

Exercise 3.4.4

Show that $\{\mathbf{v}_1 + W, \ldots, \mathbf{v}_{m-n} + W\}$ is a basis for V/W, by showing that the basis spans the space and that is linearly independent.

□

§3.5. Linear Transformations and the Rank Nullity Theorem

Linear transformations or **homomorphisms** between linear transformations is very similar to the idea of homomorphisms for groups: a linear transformation is a map T from one vector space V to another vector space W (both over the same field F), such that doing something in V and then mapping to W is the same thing as mapping to W first and then doing the things we did before. The conditions that give us this idea are (hopefully) unsurprisingly that for all $\mathbf{v}, \mathbf{v}_1, \mathbf{v}_2 \in V$ and all $a \in F$:

$$\begin{aligned} T(\mathbf{v}_1 + \mathbf{v}_2) &= T(\mathbf{v}_1) + T(\mathbf{v}_1) \\ T(a\mathbf{v}) &= (\mathbf{v}) \end{aligned}$$

In other words, T commutes with addition and scalar multiplication. And as always, isomorphisms are just bijective homomorphisms.

Gunning defines the **kernel** and **image** of a linear transformation, and we won't restate them here. Make sure to note that the kernel is a subspace of the domain of a map, and the image is a subspace of the codomain of a map (and make sure you understand why). We encourage you to try working through the (short) proof of Lemma 1.9 on your own to make sure you understand these concepts.

Exercise 3.5.1

Prove Gunning Lemma 1.9: A linear transformation $T : V \to W$ is injective if and only if $\ker(T) = \mathbf{0}$. This is actually true not just for linear transformations, but for all homomorphisms (groups, rings, fields, etc).

This next theorem is huge, that shows up not just in the context of this class, but a lot in abstract algebra.

Theorem 3.5.2

The First Isomorphism Theorem for Vector Spaces: If $T : V \to W$ is a linear transformation between finite dimensional vector spaces, then $V \ker(T) \cong T(V)$

This statement is one of those statements that is very confusing at first, but once you get an intuition for this statement, it makes a ridiculous amount of sense.
Completely informally, here's the idea of the fist isomorphism theorem. We have a linear transformation $T : V \to W$, and let $K = \ker(T)$. Since K is a subspace of v, we can take V/K to be a vector space with elements being equivalence classes of the form $\mathbf{v} + K$ for $\mathbf{v} \in V$. Lets see what happens when we throw elements of each equivalence class into T. For any element $\mathbf{v}_1 + \mathbf{k}$ in the equivalence class $\mathbf{v}_1 + K$ (where $\mathbf{v}_1$ is fixed but $\mathbf{k}$ is any vector in K), we note that because k is in the kernel of T:

$$T(\mathbf{v}_1 + \mathbf{k}) = T(\mathbf{v}_1) + T(\mathbf{k}) = T(\mathbf{v}_1)$$

This tells us that through T, all elements of an equivalence class $\mathbf{v}+K$ get mapped to $T(\mathbf{v})$.

In the quotient group V/K, addition of two equivalence classes $\mathbf{v}_1 + K$ and $\mathbf{v}_2 + K$ is just $\mathbf{v}_1 + \mathbf{v}_2 + K$, and scalar multiplying an equivalence class is just $c(\mathbf{v} + K) = c\mathbf{v}_1 + K$ (these are true since K is closed under addition and scalar multiplication). So working with V/K looks exactly like working with some subspace $\text{span}(\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ of V, just with a "$+K$" tacked on at the end of every element for purely symbolic reasons. But notice that the way V/K is acting is **exactly** the way $T(V)$ is acting. The map T sends all elements of V inside an equivalence class in V/K to some representative vector in $T(V)$, essentially collapsing each equivalence class down to a single vector. The act of quotienting V by K collapses V into a set of equivalence classes, but operations on these equivalence classes look **exactly the same** as operations of the representative vector of this equivalence class in $T(V)$, because of the structure linear transformations force on maps. So the action of mapping through T and quotienting by K are essentially the same, so V/K looks and acts exactly like $T(V)$, and the two spaces are isomorphic.

Looking at this more rigorously might help too, so let's work through the proof of this:

Proof. Consider the map $T^* : V/K \to W$ induced by the map T, meaning that an equivalence class $\mathbf{v} + K$ is mapped by T^* to whatever element T would map a representative of this equivalence class to (and this is clearly well defined by the last couple paragraphs). We've quotientied out the kernel of V, so the only equivalence class in V/K that gets mapped to the zero vector in W through T^* is the equivalence class K, which acts as the zero vector in V/K. Therefore, $\ker(T^*) = \mathbf{0}$ and T^* is injective.

Notice that when we send an equivalence class $\mathbf{v}+K$ through T^*, it collapses down to just $T(\mathbf{v})$ (or equivalently, all elements of an equivalence class map to the same element under T), so the image of V/K through T^* is the same as the image of V through T. Therefore, T^* surjects onto $T(V)$, so we have a bijective linear transformation $T^* : V/K \to T(V)$. A bijective linear transformation is just an isomorphism and we're done! □

And finally, here's a commutative diagram to help visualize what is going on (see the section on homomorphisms of groups for an explanation of what a commutative diagram is).

$$\begin{tikzcd} V \arrow[r, "T"] \arrow[d] & W \\ V/K \arrow[r, "T^*"] & T(V) \arrow[u] \end{tikzcd}$$

By applying Gunning Theorem 1.8 to the First Isomorphism Theorem, we get the last part of Gunning Theorem 1.10:

Corollary 3.5.3

The Rank Nullity Theorem: $\dim V = \dim\ker V + \dim T(V)$ for finite dimensional V and a linear transformation T.

§3.6. The Mapping Theorem and The Transition to Matrices

The Mapping theorem essentially tells us that to study a linear transformation $T : V \to W$, its enough to just look at what T does to a basis of V. This will be extremely helpful for us to transition from the abstract vector spaces we've been working with to matrices.

Theorem 3.6.1

The Mapping Theorem (Gunning theorem 1.11): Let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be a basis for V.

- $T : V \to W$ is completely determined by $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$
- T is injective if and only if $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$ are linearly independent
- T is surjective if and only if $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$ span W
- Two vector spaces V and W are isomorphic if and only if they have the same dimension.

The proof of these theorems aren't hard, but doing them will make sure you have a good understanding of all of the concepts we've talked about so far. We recommend working through them yourself if you have time. Also, try this quick exercise:

Exercise 3.6.2

Use the mapping theorem to show that the image of a linear transformation cannot have a larger dimension than it's domain (i.e. $\dim V \geq \dim T(V)$).

Gunning briefly mentions that the set of all linear maps from spaces V to W form a vector space, and proves that a linear transformation has an inverse if and only if it is an isomorphism between spaces. There's nothing very enlightening to say about these topics, so we won't discuss them further.

Instead, we will begin to discuss how we can relate abstract vector spaces to concrete coordinate systems, in order to shift our study of linear algebra toward matrices.

Definition 3.6.3

We define a **Kronecker vector** δ_i of a vector space F^n to vector with a 1 in the ith coordinate and a 0 in all other coordinates. For example, in $\mathbb{R}^4$:

$$\delta_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \delta_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix}, \delta_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}, \delta_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

The great thing about Kronecker vectors is that (1) the n Kronecker vectors for a space F^n form a basis for F^n and (2) they are very nice to work with computationally. The mapping theorem tells us that any n-dimensional vector space V over a field F is isomorphic to F^n, just by sending each basis vector in a basis of V to a distinct δ_i in F^n. This gives us a bijection between bases and isomorphisms.

$$(\text{choice of ordered basis for } V) \Leftrightarrow (\text{isomorphism } V \to F^n)$$

What this allows us to do is view each basis of V as a **coordinate system** of F^n, by associating each basis vector of V to a "direction" in F^n. Isomorphisms from one vector space to another is just the formalization of the idea of switching coordinate systems in a space.

Given a linear transformation A from an m dimensional vector space to an n dimensional vector space W, we can build a map from F^m to F^n:

$$\begin{tikzcd} V \arrow[r, "A"] & W \arrow[d, "T(\cong)"] \\ F^m \arrow[u, "S(\cong)"] \arrow[r, "B"] & F^n \end{tikzcd}$$

The mapping B we've build is just the mapping TAS. What we've done here is we've represented a mapping of abstract vector spaces V and W to a mapping of concrete vector spaces F^m and F^n, that other than changing coordinate systems, preserves the information in the "abstract world." This is great, since we can now study linear algebra just by working with these easier, concrete spaces.

§3.7. Matrices

Matrices just happen to be a very convenient way to format the information of where a basis gets sent in a mapping from F^m to F^n. By the mapping theorem, this mapping T from F^m to F^n is fully determined by where the vectors $\delta_1, \ldots, \delta_m$ in F^m get sent. A basis vector δ_j in F^m will be sent to $T(\delta_j) = \sum_{i=1}^n a_i \delta_i$ in F^n. The coefficients a_{ij} (the coefficient on the δ_i coordinate of the image of δ_j) completely define the mapping T. An easier way to think about this is that the coefficients down the jth column are the coefficients $\sum_{i=1}^n a_i \delta_i$ for the image of δ_j.

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ a_{31} & a_{32} & \ldots & a_{3m} \\ & & \ldots & \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix} = \begin{pmatrix} | & | & & | \\ T(\delta_1) & T(\delta_1) & \ldots & T(\delta_m) \\ | & | & & | \end{pmatrix}$$

Since this matrix represents a transformation from F^m to F^n, we end up with m columns (since there are m basis vectors to in F^m) and n rows (since there are n basis vectors that represent the images of the basis vectors in F^m), so we get a $n \times m$ matrix (it's customary to denote the size of a matrix as ”row” $\times$ ”column”.

Exercise 3.7.1

Argue why the identity matrix, a square matrix with 1s down the diagonal and 0s everywhere else, acts as an identity transformation. For a $m \times m$ identity matrix, where does each basis vector in F^m get sent to?

We would like to define multiplication between a vector $\mathbf{v}$ in F^m and an $n \times m$ matrix representing a linear transformation $T : F^m \to F^n$ in a way that returns the image of $\mathbf{v}$ in F^n. We know that what we want needs to look something like the definition of how vectors go through linear transformations:

$$T\begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} = T(v_1\delta_1 + \cdots + v_n\delta_n) = v_1T(\delta_1) + \cdots + v_mT(\delta_m)$$

Numerically, this will look like multiplying the ith row of T with the column vector $\mathbf{v}$ coordinat e wise and adding all the terms together, like a dot product (convince yourself why the dimensionality of this works out).

$$\begin{pmatrix} a_{11} & a_{12} & \dots & a_{1m} \\ a_{21} & a_{22} & \dots & a_{2m} \\ a_{31} & a_{32} & \dots & a_{3m} \\ & & \dots & \\ a_{n1} & a_{n2} & \dots & a_{nm} \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_m \end{pmatrix} = \begin{pmatrix} a_{11}v_1 + a_{12}v_2 + \cdots + a_{1m}v_m \\ a_{21}v_1 + a_{22}v_2 + \cdots + a_{2m}v_m \\ a_{31}v_1 + a_{32}v_2 + \cdots + a_{3m}v_m \\ \dots \\ a_{n1}v_1 + a_{n2}v_2 + \cdots + a_{nm}v_m \end{pmatrix}$$

Exercise 3.7.2

Compare this ”row-column multiplication” to the definition of how vectors go through linear transformations above to convince yourself why this numerically works.

Similarly, we define matrix multiplication in a way that captures the idea of composing two transformations:

$$\underbrace{\begin{pmatrix} & & \\ & A & \\ & & \end{pmatrix}}_{\textbf{A}} \underbrace{\begin{pmatrix} | & | & & | \\ \mathbf{v}_1 & \mathbf{v}_2 & \dots & \mathbf{v}_l \\ | & | & & | \end{pmatrix}}_{\textbf{B}} = \underbrace{\begin{pmatrix} | & | & & | \\ A\mathbf{v}_1 & A\mathbf{v}_2 & \dots & A\mathbf{v}_l \\ | & | & & | \end{pmatrix}}_{\textbf{AB}}$$

What is going on here is that B is transforming basis vectors in F^l to it's codomain, say F^m, and then these vectors are being transformed by A, leaving us with a matrix describing where the basis vectors in F^l go after being operated on by the composition AB.

Note that the multiplication of matrices with matrices and matrices with vectors needs specific dimensions to match up in order for multiplication to be allowed. We multiply matrices from right to left (since they act like the composition of mappings), so AB can only be a well-defined composition if the dimension of the codomain of B is equal to the domain of A, or in other words, the number of rows of B must match the number of columns of A. The same logic works for multiplying matrices with vectors: the product $A\mathbf{v}$ only makes sense if the domain of A is the same dimension as the vector $\mathbf{v}$, or in other words, the number of columns of A must match the number of rows in $\mathbf{v}$.

To understand the concepts of multiplying vectors in matrices in a much more visual and animated way, we highly recommend watching Chapters 3 and 4 3Blue1Brown's Essence of Linear Algebra series on YouTube (each video is about 10 minutes long). As a final exercise for this section, try "inventing" the idea of a change of basis matrix:

Exercise 3.7.3 (Insert change of basis exercise from google form)

§3.8. Equivalent Matrices and the Equivalence Theorem

We want to put an equivalence relation on matrices that group matrices by their what they do, regardless of the coordinate system that it's done in. With this motivation, we define an equivalence relation $A \asymp B$ whemever $B = SAT$ for invertible (square) matrices S and T. The idea behind this is that the matrix T is an isomorphism mapping B to the coordinate system that A acts on, then A performs the action in B but in the current coordinate system, and then S maps the image of A back into the coordinate system of B. What we'd like to explore is if there is a "nice" representative in each equivalence classes.

Theorem 3.8.1

Slightly more general Gunning Theorem 1.15: Given a map $A : V \to W$ where $\dim V = m$ and $\dim W = n$, there exist isomorphisms $S : V \to F^m$ and $T : W \to F^n$ such that

$$TAS = \begin{pmatrix} I_r & 0 \\ 0 & 0 \end{pmatrix}$$

Where $r = \dim A(V)$

Proof. A commutative diagram helps see how to approach this proof:

$$\begin{tikzcd} V \arrow[r, "A"] & W \arrow[d, "T(\cong)"] \\ F^m \arrow[u, "S(\cong)"] \arrow[r, "B"] & F^n \end{tikzcd}$$

All this proof consists of is "building" this map B by constructing S and T isomorphisms that do what we want. Looking at the form we would like TAS to be in, we want B to send the vectors δ_i in F^m for $1 \le i \le r$ to themselves in F^n, and B to send the vectors δ_i in F^m for $r < i \le m$ to the zero vector in F^n. We would like to do all of this through A. Since isomorphisms have a trivial kernel (the zero vector in the domain is the only vector sent to the zero vector in the codomain), A needs to be the transformation that "kills off" all the δ_is for $i > r$. So we would like S to send all the δ_is for $i > r$ to the kernel of A, and the rest of the δ_is to a non-kernel part of A. Since we construct isomorphisms by sending bases to bases (by the mapping theorem), this motivates us to construct a basis for A by building off a basis for $\ker A$, in order to be able to work with $\ker A$ disjointly from the rest of A.

We'll sketch out the formality of this proof, and leave it to you to fill in the details. By rank-nullity, the dimension of $\ker A$ is $m - r$, so we define a basis $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_m$ on $\ker A$ (we have $m - r$ vectors in this basis). In the same way as Gunning Theorem 1.8, we build the rest of the basis of A from $\ker A$ by adjoining linear independent vectors until we reach $\dim A$ basis vectors. This gives us a basis $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of A. Through A, the vectors $\mathbf{v}_{r+1}, \ldots, \mathbf{v}_m$ go to zero (since they're in the kernel), and r vectors $\mathbf{v}_1, \ldots, \mathbf{v}_r$ form a basis for $A(V)$ (again by mapping theorem).

To send all the δ_is in F^m for $i > r$ to zero in F^n, we have S send δ_i to $\mathbf{v}_i$ for $i > r$. Pushing these $\mathbf{v}_i$s through A kills them to zero, and they stay zero through T. To send all the δ_is in F^m for $i \le r$ to themselves in F^n, we have S send δ_i to $\mathbf{v}_i$ for $i \le r$, and A sends them to a basis $T(\mathbf{v}_1), \ldots, T(\mathbf{v}_n)$ that spans $T(V)$. We then just define T to be an isomorphism that sends $T(\mathbf{v}_i)$ to δ_i for $ileqr$, and we have the desired transformation that produces the desired matrix. Here is a more detailed commutative diagram to help recap the ideas of the proof.

$$\begin{tikzcd}
v_i & V \arrow[r, "A"] & W \arrow[d, "T(\cong)"] & w_i \arrow[d] & 0 \arrow[d] \\
\delta_i \arrow[u] & F^m \arrow[u, "S(\cong)"] \arrow[r, "B"] & F^n & \delta_i & 0
\end{tikzcd}$$

□

Still need to do rank stuff but I don't feel like doing it tn this is a tomorrow problem.

§3.9. Alternating Multilinear Forms and the Determinant

- Define a transposition - Define sign - Show sign is well defined.

§3.10. Key Takeaways

§3.11. A Bunch of Definitions

$$A = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1m} \\ a_{21} & a_{22} & \ldots & a_{2m} \\ a_{31} & a_{32} & \ldots & a_{3m} \\ & & \ldots & \\ a_{n1} & a_{n2} & \ldots & a_{nm} \end{pmatrix} = \begin{pmatrix} \\ T(\delta_1) & T(\delta_1) & \ldots & T(\delta_m) \\ \\ \end{pmatrix}$$

beginalign* $\begin{pmatrix} \\ A \\ \\ \end{pmatrix} \begin{pmatrix} \\ \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_l \\ \\ \end{pmatrix} = \begin{pmatrix} \\ A\mathbf{v}_1 & A\mathbf{v}_2 & \ldots & A\mathbf{v}_l \\ \\ \end{pmatrix}$

5. Metric Spaces

A lot of understanding metrics comes from learning the ways the tools of metrics can be used in proofs, so in this chapter we'll be spending a lot of time focusing on the intuition, motivation, and understanding behind proofs, and not spend too much time on definitions that Gunning discsusses in his book.

§5.1. Defining Metrics

While a norm is a sense of the "length" of a vector, a metric is a sense of "distance" between two points. This is the basis for the intuition of pretty much everything with metrics.

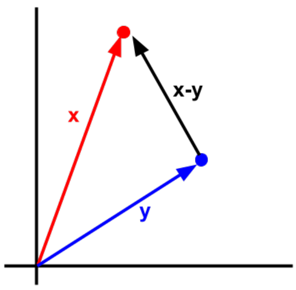
Definition 5.1.1

We define a **metric** ρ on a set S as a mapping between coordinates in $S \times S$ to the real numbers (technically it can be to any ordered field but we'll just stick with the reals) which has the following properties:

- Positivity: $\rho(x, y) \geq 0$ and $\rho(x, y) = 0$ iff $x = y$
- Symmetry: $\rho(x, y) = \rho(y, x)$
- The Triangle Inequality: $\rho(x, y) = \rho(x, z) + \rho(y, z)$.

It makes intuitive sense why positivity and symmetry would be included in the definition of a metric, and the triangle inequality plays the same role as it does in the definition of norms. [TO DO: EXPLAIN TRIANGLE INEQUALITY WELL IN NORMS].

A metric space is less restricted than a norm, since while norms operate on a vector space, metrics operate on sets. Because of this heirarchy, any norm is able to induce a metric on a vector space (where the vector space is viewed as just a set). The notion of distance in this norm-induced metric is the length of the difference of two vectors, $\rho(x, y) = ||x - y||$ (where x and y are vectors). This can be seen more clearly by the picture below:



One extremely important metric is the **discrete metric**, which is a metric ρ on a set S where $\rho(x, y) = 1$ whenever $x \neq y$, and $\rho(x, y) = 0$ when $x = y$. You can check that this satisfies the properties of a metric. This means that every point in S is an equal distance away from every other point in S.

Gunning defines metric subsets and denseness in this section, but there isn't anything particularly enlightening about it so say here. One example to keep in mind is that $\mathbb{Q}$ is dense in $\mathbb{R}$ under the metric induced by the l_1 (or absolute value) norm.

§5.2. Pseudometrics, Equivalent Metrics, and Isometries

Sometimes, we can have a mapping ρ' from a set S to the reals that satisfies all of the metric axioms, except that $\rho'(x,y) = 0$ for some x, y where $x \neq y$. This is what we call a **pseudometric**. In a normal metric, $\rho'(x,y) = 0$ should directly imply that $x = y$, but here there is ambiguity. As usual, we can use the tool of quotients to "glue" the ambiguities together, forcing $x = y$ whenever $\rho'(x,y) = 0$, and gluing other parts of the space together to allow this to happen while still preserving the structure given by the second and third axioms of a metric space. Rigorously, we define an equivalence relation $x \asymp y$ whenever $\rho'(x,y) = 0$, and quotient S by this equivalence relation. To make sure that this quotient respects the metric properties, we need to show that our metric is well defined on equivalence classes.

Exercise 5.2.1

Show that ρ' is well defined on equivalence classes, i.e. if $x \asymp x'$ and $y \asymp y'$, then $\rho(x,y) = \rho(x',y')$.

This process allows us to turn any pseudometric ρ' on S to a metric ρ' on $S/\asymp$ (note: we're quotienting the space, but we're not actually touching the function that ρ defines).

Equivalent metrics work the exact same way equivalent norms do, and equivalent norms define equivalent metrics.

A good way to think about isometric embeddings are like homomorphisms of groups. Just like we would like a homomorphism ϕ between groups G and H to preserve the group structure of both of them, we would like an isometric embedding ψ between spaces S (with metric ρ_S) and T (with metric ρ_T) to preserve the structure of a metric. This motivates us to define an isometric embedding as a mapping that satisfies the following restriction:

$$\text{For all } x, y \in S, \rho_S(x,y) = \rho_T(\psi(x), \psi(y))$$

Essentially, what this says is that taking the "distance between x and y in S" is the exact same as mapping x and y into T through ψ, and then "taking the distance in T." Just like groups, visualizing an isometric embedding through a commutative diagram may help. Let x and y be points in S, and let ψ be an isometric embedding between spaces S (with metric ρ_S) and T (with metric ρ_T). Then we have:

$$\begin{tikzcd}
(x,y) \arrow[r, "\psi"] \arrow[d, "\rho_S"] & (\psi(x), \psi(y)) \arrow[d, "\rho_T"] \\
\rho_S(x,y) \arrow[r, "\psi"] & \rho_T(\psi(x), \psi(y))
\end{tikzcd}$$

Exercise 5.2.2

Show that an isometric embedding must be injective.

Isometries are defined as bijective isometric embeddings, and they are analogous to isomorphisms in the sense that isometric metric spaces are essentially the same metric space and act in the same ways, just like how isomorphic groups are essentially the same and act in the same ways.

§5.3. Sequences and Cauchy vs Convergent

We define a **sequence** as a set S indexed by the natural numbers. We can denote a sequence $\{a_1, a_2, \dots\}$ as $(a_n)_{n=1}^{\infty}$, or more simply, (a_n). A sequence converges to some value a in S with respect to a metric ρ if it gets arbitrarily close to this a as n gets very large, or more rigorously:

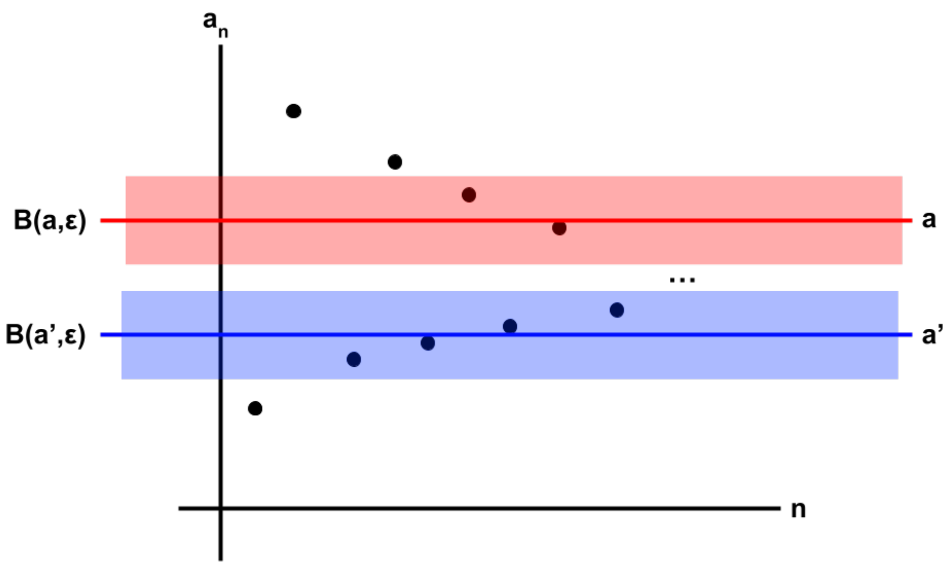
For all $\varepsilon > 0$ there exists a natural number N such that for all $n > N, \rho(a, a_n) < \varepsilon$

If a sequence (a_n) converges to a value a, we say that $lim_{n\to\infty} a_n = a$. To practice using these definitions, let's prove the following lemma:

Lemma 5.3.1

Let (S, ρ) be a metric space. If a sequence (a_n) converges to both a and a', then $a = a'$. In other words, limits are unique in metric spaces (this isn't necessarily true for topological spaces, as we'll see).

Proof. Your best friend in all proofs in this chapter will be the triangle inequality and drawing pictures. So let's do both of them here.



We draw a picture to build intuition. Let's assume that there are two values a and a' that our sequence converges to. By the definition of convergence, all elements of (a_n) past a certain point will fall a distance ε from a and a'. We draw a ball of radius ε around a, denoted $B(a, \varepsilon)$ (colored red in the picture), and do the same for a' (colored blue in the picture). We notice that if we make epsilon "small enough", then the two balls of radius epsilon become disjoint. This leads to a contradiction, since we can't have points be in both of these balls at the same time. With this intuition in mind, we construct a rigorous argument.

We pick an ε such that $\epsilon < \frac{\rho(a,a')}{2}$ (this forces the balls to be disjoint). Then by the definitions of convergence, we have an N_1 such that for $n > N_1$, $\rho(a_n, a) < \varepsilon$, and we have a N_2 such that for $n > N_2$, $\rho(a_n, a') < \varepsilon$. We let N be the maximum of N_1 and N_2. For $n > N$, by triangle inequality, we get that $\rho(a, a') \leq \rho(a, a_n) + \rho(a_n, a') < \frac{\rho(a,a')}{2} + \frac{\rho(a,a')}{2} = 2\varepsilon$, a contradiction since $\rho(a, a') = 2\varepsilon$. $\square$

This technique of drawing pictures to build intuition and then formalizing an argument of distance using triangle inequality shows up over and over again, so keep this example in mind.

If ρ_1 and ρ_2 are equivalent metrics, then a sequence converges with respect to ρ_1 if and only if it converges with respect to ρ_2. This is because each metric differs by the other by at most a constant factor, so the only thing that changes is the "rate" at which the sequence converges. As an exercise to practice being comfortable with the definitions of equivalence and convergence, show this formally.

Exercise 5.3.2

Show formally that if ρ_1 and ρ_2 are equivalent metrics on a set S, then a sequence (a_n) in S converges with respect to ρ_1 if and only if it converges with respect to ρ_2.

The hard thing with determining if a sequence converges is that a lot of times, we don't really know what a sequence converges to. (I need to fix this sentence) However, we can characterize sequences that "seem" to converge to something (even if they don't), sequences where terms get closer and closer to each other as you go further and further in the sequence. We define a **Cauchy sequence** to be a sequence where going far enough in the sequence garuntees that terms get arbitrarily close to each other. Rigorously,

For all $\varepsilon > 0$ there exists a natural number N such that for all $m, n > N, \rho(a_m, a_n) < \varepsilon$

Notice the distinction between Cauchy and convergent sequences. Cauchy sequences talk about values in a sequence getting close to other values in a sequence, while convergent sequences talk about values in a sequence getting close to a fixed value not necessarily in the sequence. It's not very hard to show that all convergent sequences are Cauchy, it's just a quick application of triangle inequality.

Exercise 5.3.3

Show that a convergent sequence is necessarily Cauchy.

However, and this is extremely important, **a Cauchy sequence is not necessarily convergent.** Gunning gives a couple examples which we recommend looking at, but we'll give some here too. Consider $S = (0, 1]$ with the standard absolute value metric. It's clear that the sequence $(\frac{1}{n})_n = \{1, \frac{1}{2}, \frac{1}{3}, \ldots\}$ is Cauchy. However, it converges to 0, which is not contained in S, so this sequence is not convergent in S. The idea behind Cauchy not implying convergent is that sets are not necessarily closed under convergence, or put another way, sequences that are fully contained in a set might converge to a point outside a set. Another example is the set $S = \mathbb{Q}$ and the sequence $\{3, 3.1, 3.14, 3.141, 3.1415, \ldots\}$, which clearly converges to π, which is not in the set of rational numbers.

We call a metric space S **complete** (or Cauchy complete) if every Cauchy sequence in S converges in S. Put another way, S is closed under convergence.

§5.4. Some Limit Properties

Gunning lists some limit properties that hold in $\mathbb{R}$ but doesn't highlight them at all. However, these properties are extremely useful to know, so we'll go through them here

- $\lim_{n\to\infty}(a_n + b_n) = \lim_{n\to\infty}(a_n) + \lim_{n\to\infty}(b_n)$, provided the limits of a_n and b_n exist.
  - Proof outline: if we know the limits of a_n and b_n exist, (let's call them a and b respectively), we know that a_n and b_n converge to their limits (by the definition of a limit). We would like to show that the sequence $(a_n + b_n)$ converges to $a + b$. By triangle inequality, $|a_n + b_n - a - b| \leq |a - a_n| + |b - b_n|$, and both terms get arbitrarily small.
- If $a_n \leq b_n$ for all n then $\lim_{n\to\infty}(a_n) \leq \lim_{n\to\infty}(b_n)$.
  - Proof outline: again, let's let $\lim_{n\to\infty}(a_n) = a$ and $\lim_{n\to\infty}(b_n) = b$. If $a > b$ then we can make a ball of radius $\frac{a-b}{2}$ around a and b, and if we go far enough in both a_n and b_n, all elements of a_n will fall in our ball around a, and all elements of b_n will fall in our ball around b. However, the two balls are disjoint, and all points in the ball around a are greater than all points in the ball around b, contradicting the fact that $a_n \leq b_n$ for all n.

Important note: that these properties cannot hold for any general metric space, since a notion of addition and order isn't required for a metric space (so we couldn't have added a and b or compared the sizes of their limits). Furthermore, we need addition to act continuously (which we'll discuss rigorously in Chapter 3) so that limits play nicely with it. However, Gunning chooses to use it freely on the metric spaces he defines, so for the rest of this section, just assume that all sets S with a metric ρ have a total order and have a continuously defined operation of addition.

§5.5. Theorem 2.5 and Binary Search

A common theme that will show up in proofs is that you need to "look for something." This could be anything from "the thing that a Cauchy sequence possibly converges to," "the thing that might be the supremum to a set", "the open covering that might not have a finite open subcover" (compact sets, see Chapter 2.4), etc. There are two techniques that you'll see throughout this class to tackle these kinds of proofs: **binary search** and **falling knife**. We'll discuss the binary search technique now, and introduce the falling knife technique in Section 2.3, but once you learn the falling knife technique come back and try to prove Thm 2.5 with this technique.

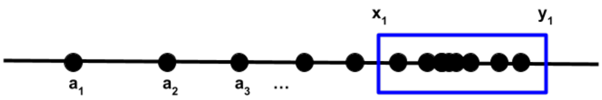
Theorem 5.5.1

Gunning 2.5: An ordered archimedian field is a complete ordered field if and only if it is a complete metric space in the absolute value metric. In other words, the notion of completeness in Section 1.2 is equivelent to the notion of completeness in this section: if F is an ordered archimedian field,

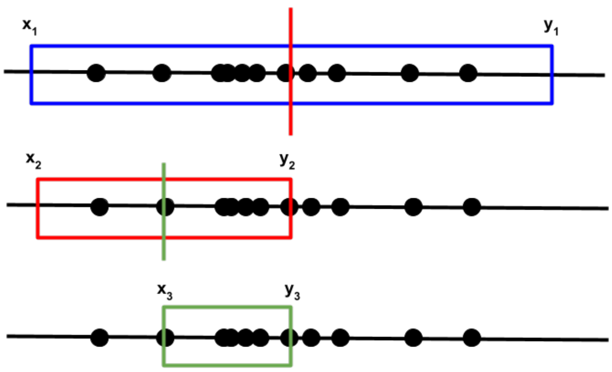
$$F \text{ has the least upper bound property } \Leftrightarrow F \text{ cauchy complete}$$

Proof. We'll start by assuming that F has the least upper bound property, which implies the existance of supremums and infimums of sets (when an upper and lower bound exist, respectively). We need to show that F is cauchy complete with respect to the absolute value metric, so let's take an arbitrary cauchy sequence (a_n) and see what we can do with it. To show that (a_n) is convergent, we need to find a value that (a_n) might converge to. So let's go searching for this value.

Since (a_n) is Cauchy, points in (a_n) will get arbitrarily close to each other as we go out far enough. Therefore, we can find real numbers x_1 and y_1 such that $x_1 < y_1$, $y_1 - x_1 \leq 1$, and there are infinitely many terms of (a_n) in $[x_1, y_1]$. You can find this x_1 and y_1 by going far enough out in the Cauchy sequence such that for $m, n > N, |a_m - a_n| < \frac{1}{2}$, so setting n to $N+1$, we get that $x_1 = a_{N+1} - \frac{1}{2}$ and $y_1 = a_{N+1} + \frac{1}{2}$ satisfies our conditions.



Now we have a region $[x_1, y_1]$ with an infinite number of a_n terms in it. If we cut this region in half (into $[x_1, \frac{x_1+y_1}{2}]$ and $[\frac{x_1+y_1}{2}, y_1]$, at least one half must have an infinite number of a_n terms in it. Call this half $[x_2, y_2]$. Intuitively, if (a_n) converges, the thing it converges to must be in the half with an infinite number of points.



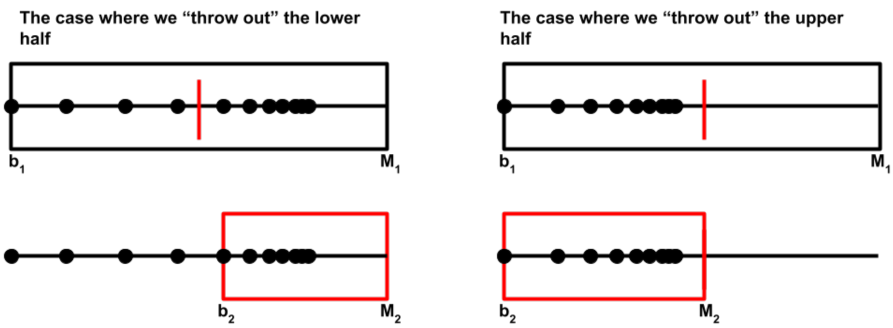
We'll continue this process, cutting $[x_2, y_2]$ in half and keeping a half of it that contains an infinite number of a_n terms, calling it $[x_3, y_3]$. Eventually we'll have an infinite number of these sets $[x_n, y_n]$, where

$$[x_1, y_1] \supset [x_2, y_2] \supset \dots \text{ with } y_i - x_i = \frac{y_1 - x_1}{2^{i-1}}$$

The $\frac{y_1 - x_1}{2^{i-1}}$ comes from the fact that we're dividing the size of $[x_i, y_i]$ in half each time. Again, our intuition tells us that if (a_n) converges, the thing it converges to should be contained in every $[x_n, y_n]$. This means that our guess of what (a_n) converges to, let's call it a, is an upper bound of the set $\{x_1, x_2, \dots\}$ and a lower bound of the set $\{y_1, y_2, \dots\}$. This is great, since we can now use the least upper bound property to say that $x_\infty = \sup x_i$ and $y_\infty = \inf y_i$ exist. We would like to show that $a = x_\infty = y_\infty$, and that a_n converges to a. There are a couple of things we need to rigorously show to complete this proof:

- $\lim_{i\to\infty} x_i = \sup x_i$ and $\lim_{i\to\infty} y_i = \inf y_i$
  - By the definition of a supremum, the x_is must get arbitrarily close to x_∞ as i gets large. But since the sequence of x_is is monotonically increasing and is bounded above by x_∞, the distance $|x_i - x_\infty|$ must be monotonically decreasing. So we get that for all $\varepsilon > 0$ there exists an N such that for $i > N$, $|x_i - x_\infty| < \varepsilon$, which tells us exactly that $\lim_{i\to\infty} x_i = \sup x_i$. A similar argument can be used to show that $\lim_{i\to\infty} y_i = \inf y_i$. Keep this idea in mind: supremum plus monotonically increasing equals limit.
- $\lim_{i\to\infty}(y_i - x_i) = 0$. Since we know $\lim_{i\to\infty}$ and $\lim_{i\to\infty}$ exist from the last bullet point, we can split $\lim_{i\to\infty}(y_i - x_i) = 0$ up into two limits to show that $x_\infty = y_\infty$
  - Here we use the Archimedian property of the **reals**. We would like to show that for any $\varepsilon > 0$ there exists an N such that for $i > N$, $|y_i - x_i| < \varepsilon$. We know that $y_i - x_i = \frac{y_1 - x_1}{2^{i-1}}$. By the Archimedian property, there exists a rational number q such that $0 < q < \varepsilon$, and since the sequence $(\frac{y_1 - x_1}{2^{n-1}})$ converges to zero, we can pick N large enough so that for $i > N$, $|\frac{y_1 - x_1}{2^{n-1}}| < q$. It's extremely important to understand that we didn't use the Archimedian property of F here, we used the Archimedian property of the reals, since the output of our limit is the range of the absolute value metric (which is the real numbers, not F).
- a_n converges to x_∞ (which is the same as converging to a since they're equal)
  - Our goal is to show that $|a_i - x_\infty|$ gets arbitrarily small as i gets large. We know how the x_is act with respect to x_∞, and we know how the a_is act with respect to x_is, so this is a hint to use triangle inequality. Note that $|a_i - x_\infty| \le |a_i - x_i| - |x_i - x_\infty|$. But since $|a_i - x_i| \le |y_i - x_i|$, and both $|y_i - x_i|$ and $|x_i - x_\infty|$ become arbitrarily small as i becomes large.

And that's the first half of the proof! Now we just have to show the other direction, that if F is Cauchy complete, F has the least upper bound property. We start with some arbitrary set S of elements of F that is bounded above by some M_1, and we'd like to show that there exists a least upper bound of S. The idea is that we are going to binary search again for a guess M of the least upper bound, and then prove that our guess is right. Let's start with the interval $[b_1, M_1]$ where b is some element of S, and clearly if a least upper bound exists it must be in this interval, since the least upper bound cannot be less than b_1. We split this interval in two halves $[b_1, \frac{M_1+b_1}{2}]$ and $[\frac{M_1+b_1}{2}, M_1]$. If there are no elements of S in $[\frac{M+b}{2}, M]$ (the upper half), set $[b_2, M_2] = [b_1, \frac{M_1+b_1}{2}]$ (the lower half). If there is an element b_2 of S in $[\frac{M_1+b_1}{2}, M_1]$ (the upper half), set $[b_2, M_2] = [\frac{M_1+b_1}{2}, M_1]$ (the upper half). Hopefully the picture below makes this more clear.



(We would like the lower ends of our intervals to be points in S for our argument to work cleanly)

We keep doing this process, and just like before, we have that

$$[b_1, M_1] \supset [b_2, M_2] \supset \dots \text{ with } |M_i - b_i| \leq \frac{M_i - b_i}{2^{i-1}}$$

Note that we have a less than or equals sign now rather than an equality, since choosing a b_2 when we keep the upper half will cut off the interval $[M_1, b_2]$.

We now have two sequences, the b_is (a monotonically increasing sequence) and the M_is (a monotonically decreasing sequence). Showing that $\lim_{i\to\infty}(M_i - b_i) = 0$ follows similarly from earlier, where we use the Archimedian property of the reals (here it's really important we don't assume the Archimedian property of F since we're trying to prove that). However, **we don't know** as of know if the individual limits of M_i and b_i exist. This is where our assumption of completeness come in: since we know the sequence of b_is and the sequence of M_is are both Cauchy, they both converge, and so the individual limits of M_i and b_i exist. Now we can say that $\lim_{i\to\infty} M_i = \lim_{i\to\infty} b_i)$, and based off our intuition, we set our guess of a least upper bound M as this limit. We leave it to you to show that M satisfies the two criteria for a least upper bound: that M is an upper bound for S and that no value in F smaller than M is an upper bound for S.

Exercise 5.5.2

Fill in the details of the end of this proof.

□

This concept of binary searching, or more generally, "squeezing" something we want to find around two sequences that converge to the same value, is an extremely useful technique that we'll see throughout the rest of this chapter and into Chapter 3 (the squeeze theorem uses a very similar concept).

ADD SOMETHING HERE ABOUT EQUIVALENT NORMS

§5.6. Equivalent Cauchy Sequences and the Very Important Theorem 2.7

Before we approach Theorem 2.7, we need to discuss the concept of equivalent Cauchy sequences. We can define an equivalence relation $\asymp$ on Cauchy sequences so that $(a_n) \asymp (b_n)$, for any $\varepsilon > 0$ there exists an N such that for $m, n > N$, $\rho(a_m, b_n) < \varepsilon$ (or put more simply, the distance between a point in (a_n) and a point in (b_n) get arbitrarily close as n gets large). Let's show some properties of equivalent Cauchy sequences:

Lemma 5.6.1
- (i) Two Cauchy sequences (a_n) and (b_n) in S are equivalent if and only if $lim_{n\to\infty}\rho(a_n, b_n) = 0$
- (ii) if (a_n) and (b_n) are equivalent Cauchy sequences in S, then (a_n) converges exactly when (b_n) does, and the two sequences converge to the same value.

Proof. We'll provide an outline here, and you can fill in the details on your own if you'd like.

(i) If two Cauchy sequences (a_n) and (b_n) are equivalent, then we set $m = n$ to get that past some N, $n > N$ implies $\rho(a_n, b_n) < \varepsilon$. This gives us that for all $\varepsilon > 0$ there exists an N such that $|\rho(a_n, b_n) - 0| < \varepsilon$, implying that $\lim_{n\to\infty} \rho(a_n, b_n) = 0$. We write this last part out to emphasize that the limit of a metric is a sequence of real numbers, not a sequence of values in S, so since the real numbers are equipped with the absolute value metric, we use this to compare the "distance" between $\rho(a_n, b_n)$ and 0.

If two sequences (a_n) and (b_n) satisfy $lim_{n\to\infty}\rho(a_n, b_n) = 0$, we know that for all $\varepsilon > 0$ there exists an N such that $\rho(a_n, b_n) < \varepsilon$. By triangle inequality, we not that $\rho(a_m, b_n) \leq \rho(a_m, b_m) + \rho(b_m, b_n)$, and both of these terms become arbitrarily small as you go far enough in (a_n) and (b_n) (if you would like to make this rigorous, find an N such that both $\rho(a_m, b_m)$ and $\rho(b_m, b_n)$ are smaller than some $\frac{\varepsilon}{2}$).

(ii) Let (a_n) converge to a, then for any $\varepsilon > 0$ there exists an N such that for $n > N$, $\rho(a_n, a) < \varepsilon$. Just slap triangle inequality on $\rho(b_n, a)$ to show that it's less than $\rho(b_n, a_n) + \rho(a_n, a)$ to show that it gets arbitrarily small too. In Gunning's proof, to show that (b_n) converges to a, he shows that $lim_{n\to\infty}\rho(b_n, a) = 0$. We leave it as an exercise to show that these two notions of convergence are equivalent

Exercise 5.6.2

Show that a sequence (a_n) converging to a with respect to a metric ρ is equivalent to showing that $lim_{n\to\infty}\rho(a_n, a) = 0$.

□

Now that we have the tools of equivalent Cauchy sequences, we can start to tackle the proof of Gunning Theorem 2.7. In order to construct the real numbers from the rational numbers, we turn $\mathbb{Q}$ into a complete metric space, adjoining all of the values that Cauchy sequences in $\mathbb{Q}$ "converge to", and closing the space with respect to convergence. Theorem 2.7 generalizes this procedure.

Theorem 5.6.3

Gunning 2.7: For any real metric space S there exists a canonical isometry $\phi : S \to S^{**}$ into a complete metric space S^{**}, and the image $\phi(S) \subset S^{**}$ is a dense subspace of S^{**}

Proof. First of all, the statement of this proof is already very technical, intertwining many of the new definitions we've learned. To gain an intuition on what this statement is saying, we recommend thinking of S as $\mathbb{Q}$ and S^{**} as $\mathbb{R}$, with the isometry acting as the identity on $\mathbb{Q}$. Our isometry injects $\mathbb{Q}$ into $\mathbb{R}$ while preserving the structure of the metrics on the two sets, and the image of $\mathbb{Q}$, which is just $\mathbb{Q}$, is dense in $\mathbb{R}$.

To construct S^{**} from S (and we use our intuition going from $\mathbb{Q}$ to $\mathbb{R}$), we would like to "make all the Cauchy sequences converge." From (S, ρ), we can construct a set $S*$ of all Cauchy sequences in S (so the elements of $S*$ are Cauchy sequences), with a pseudometric ρ^* that tells us "how far" Cauchy sequences are from each other. Then, by quotienting by equivalent Cauchy sequences, Cauchy sequences that "converge" to a certain number (say, $\sqrt{2}$) can define an equivalence class representing this number. In this way, we're "completing" the set S. Now lets make this rigorous.

To define a pseudometric on the set S^* of Cauchy sequences in S, where equivalent Cauchy sequences have a distance 0 apart with respect to ρ^*, we would like to define a pseudometric ρ^* in a way that allows us to quotient by equivalent Cauchy sequences later. Since Cauchy sequences are equivalent iff $\lim_{n\to\infty} \rho(a_n, b_n) = 0$, we are motivated to define our pseudometric ρ^* as

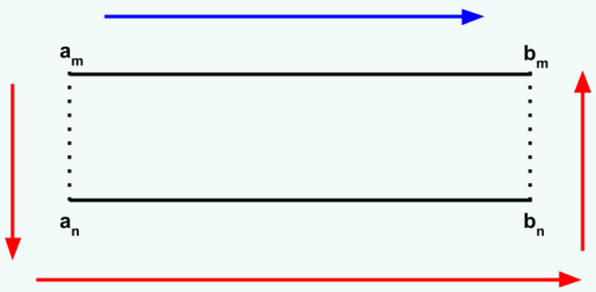
$$\rho^*((a_n), (b_n)) = \lim_{n\to\infty} \rho(a_n, b_n)$$

(Think about why this is a pseudometric and not just a metric) However, in order for ρ^* to be well defined, we need to make sure that the limit on the right side exists for all $\rho(a_n, b_n)$. Let's do this first.

Lemma 5.6.4

If (a_n) and (b_n) are Cauchy sequences in S, then the sequence $(\rho(a_n, b_n))$ converges, and therefore the limit we are concerned about always exists.

Proof: This should be screaming triangle inequality at you. It helps to draw a picture to visualize how we should apply it.



We want to show that the sequence $(\rho(a_n, b_n))$ of real numbers is Cauchy, because since the real numbers are a complete metric space (with respect to the absolute value metric), we can say that $(\rho(a_n, b_n))$ converges and our limit exists. To show this, we must show that for any $\varepsilon > 0$ there exists an N such that for $m, n > N$, $|\rho(a_m, b_m) - \rho(a_n, b_n)| < \varepsilon$. We know by triangle inequality that $\rho(a_m, b_m) \leq \rho(a_m, a_n) + \rho(a_n, b_n) + \rho(b_n, b_m)$. Rearranging, we get that $|\rho(a_m, b_m)\rho(a_m, a_n)| \leq \rho(a_n, b_n) + \rho(b_n, b_m)$ (where the absolute value comes from the fact that we can reverse the roles of n and m). Since the right hand side gets arbitrarily small, we've proved our lemma.

Now that we're sure ρ^* is well-defined, we can easily check that ρ^* satisfies the properties of a pseudometric. We can quotient out by equivalent Cauchy sequences in S^* to get a metric space (S^{**}, ρ^{**}).

Before we continue, let's define some notation. Let a capital letter A or B stand for an element in S^*, meaning they represent a Cauchy sequence (a_n) or (b_n). Let $[A]$ denote the equivalence class containing A of Cauchy sequences in S^*. This equivalence class is an element of S^{**}.

Now we need to define our isometric embedding ϕ injecting S into S^{**}, just a way to represent each element of S in S^{**}. There's a clear bijection between elements of S and constant Cauchy sequences in S^*, mapping $a \in S$ to $\{a, a, a, \dots\}$ in S^*. And since for $a \neq b$, $\{a, a, a, \dots\}$ converges to a different value that $\{b, b, b, \dots\}$, they are in different equivalence classes in S^{**} when we quotient. We'll denote the constant sequence $\{a, a, a, \dots\}$ as $C(a)$ from now on for convenience.

To complete the proof, we just need to check that S^{**} satisfies the properties stated in the Theorem:

- (1) S is dense in S^{**}
- (2) S^{**} is complete
- (3) $\phi : S \to S^{**}$ is an isometric embedding.

Lets do it! When you have many statements you need to prove, it's often easier to look at the statements and try to figure out which ones follow quickly and which ones are going to require a lot more work.

Lemma 5.6.5
(3) is true.

Proof: We need to show that for $a, b \in S, \rho(a, b) = \rho^{**}([C(a)], [C(b)])$. By the definition of how (S^{**}, ρ^{**}) is constructed from (S^*, ρ^*), we have that

$$\rho^{**}([C(a)], [C(b)]) = \rho^*(C(a), C(b))$$

And by how we defined ρ^*, we have that

$$\rho^*(C(a), C(b)) = \rho(a, b)$$

This completes the proof of (3).

Lemma 5.6.6
(1) is true.

Proof: We need to show

$$\begin{aligned} \text{For all } [A] \in S^{**}, \text{ and for all } \varepsilon > 0, \text{ there exists } x \in S \text{ such that } & \rho^{**}(\phi(x), [A]) < \varepsilon \\ \Leftrightarrow & \rho^*(C(x), [A]) < \varepsilon \\ \Leftrightarrow & \lim_{n \to \infty} \rho(x, a_n) < \varepsilon \end{aligned}$$

We know that (a_n) is a Cauchy sequence, so we know that for some $m, n > N$, $\rho(a_m, a_n) < \varepsilon$. We can fix $x \in S$ in our limit by making it independent of n, so if we fix x to be a_N, for all $n > N$, $\rho(a_n, a_N) < \varepsilon$, completing the proof of (1).

We actually have something stronger here. INSERT STRONGER THING.

Lemma 5.6.7

(2) is true.

Proof: We need to show that any Cauchy sequence of equivalence classes of elements of $S*$ (which are Cauchy sequences of S) converges. Make sure you understand this sentence before you go on.

FINISH THIS LATER

And we're done! $\square$

Although this is a very challenging proof, there isn't really a sneaky clever step that makes this proof tricky. It's the length and technicality of this proof that makes it as difficult as it is. However, every step that we made was strongly motivated, and as long as we kept our eyes on the goal of the proof, we were able to make it through it. Being able to recreate this proof without the help of this guide or Gunning is an extremely helpful exercise, both to strengthen your understanding of how to use the tools in this chapter, and to build confidence in using motivation to guide in problem solving.

Exercise 5.6.8

Recreate the proof of Gunning 2.7 on your own.

§5.7. The Bridge to Topologies

Metrics are a special case of topologies, which we'll discuss in the next chapter. To start building intuition on the ideas of topology, we'll introduce a couple topological concepts and apply them to metric spaces.

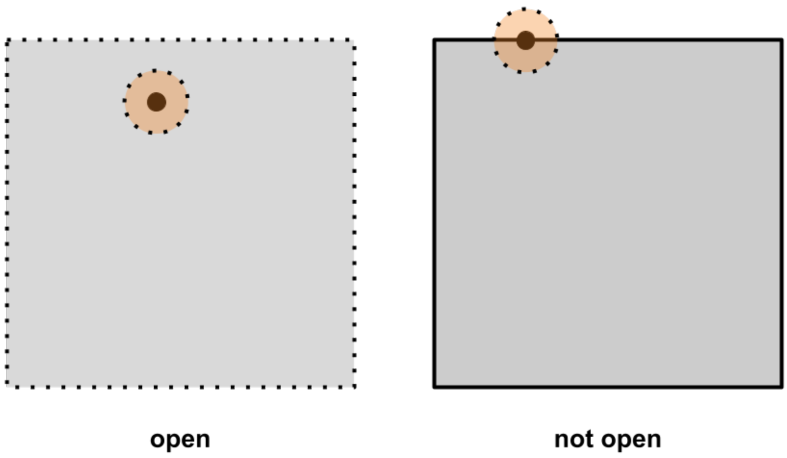
Definition 5.7.1

An **epsilon-neighborhood** of a point a in a set S, denoted $N_\varepsilon(a)$, is a ball of radius ε around the point a, not including the boundary. More rigorously, it is defined as

$$N_a(\varepsilon) = \{x \in S : \rho(a, x) < \varepsilon\}$$

An **open set** U is some subset of S such that if I pick any point a in U, I can find a neighborhood of a that is completely contained in U.

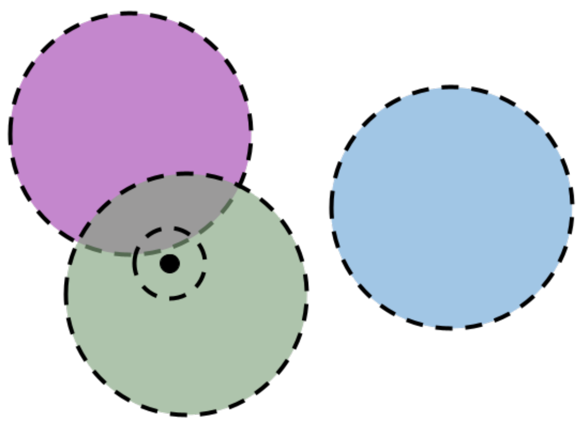
Pictures help with these kinds of concepts:



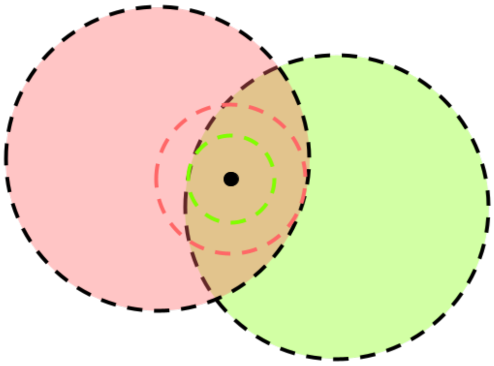
The square on the left doesn't include it's boundary, so if we put a point anywhere in the interior of the square, we can draw a little circle around it (our epsilon neighborhood) such that the circle is entirely in the interior of the square. This means the square on the left is open. The square on the right does include it's boundary, and it is not possible to make an epsilon neighborhood of a point on the boundary completely fit within the square and it's boundary. Therefore, the square with the border is not open.

Some important things to keep in mind about open sets in metric spaces (along with pictures to help) are below. In a metric space S:

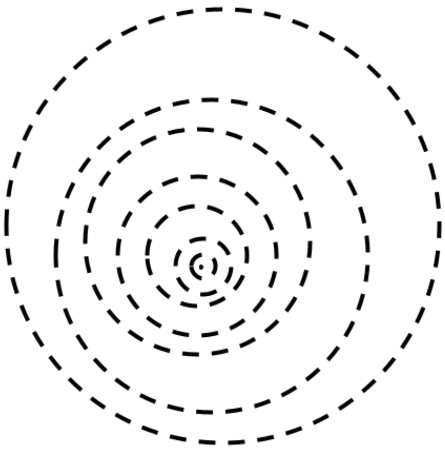
- The empty set is an open set, since there are no points to check to see if neighborhoods around these points are contained in the empty set. The entire set S is an open set, since by definition, a neighborhood of a point in S is a subset of S.
- If A and B are open sets in S, then $A \cup B$ is an open set in S. To show this, we need to show that every point in $A \cup B$ has an epsilon neighborhood completely contained in $A \cup B$. But this is already true for every point in A, since $A \subset A \cup B$, and same for B. In this way, we can argue that **the union of an arbitrary (not necessarily finite or countable) number of open sets in S is an open set**.



- If A and B are open sets in S, then $A \cap B$ is an open set in S. To show this, we need to show that every point in $A \cap B$ has an epsilon neighborhood completely contained in $A \cup B$. Let's take some generic point $x \in A \cap B$. We note that we have some neighborhood N_a of x that is completely contained in A, and some neighborhood N_b of x that is completely contained in B. Taking the smaller of the two neighborhoods gives us a neighborhood of x that is contained in $A \cap B$.



We can generalize this argument, but we have to be **very** careful. Note that in extending this argument to the intersection of more than two open sets, we are taking the minimum of the radii of the epsilon neighborhoods fitting in each of our open sets. If we have a finite number of open sets we are intersecting, we are taking the minimum of a finite number of nonzero quantities, which will always return a nonzero quantity. However, if we have an infinite number of open sets, **it is possible that the radii could form an infinite sequence of positive numbers who's infimum is zero**. An epsilon neighborhood must have a nonzero radius, so this shows that **it is true that the intersection of a finite number of open sets in S is an open set, but this statement is not necessarily true for the intersection of an infinite number of open sets**.



Keep these ideas of "open sets of S" in mind, since these ideas form the basis of point set topology, which we'll talk about in sections 2.3 and 2.4.

§5.8. Key Takeaways

This chapter introduces a lot of new concepts, and you'll get a good feeling for them through the 2.2 homework problems. However, this chapter also contained a lot of very,

very important problem techniques (both for metric problems and for math problems in general) that will continue to be extremely useful throughout 216 (and probably after this class too, we don't know yet), so let's do a quick recap:

- **Draw pictures whenever you can**, especially for metric problems. The kind of "graph and dots" drawing in Lemma 5.3.1 is extremely helpful for getting an idea on how to approach metric proofs.
- **The triangle inequality is your friend**. It will show up literally everywhere. Review some of the proofs of this chapter and note how triangle inequality showed up. When you see triangle inequality come up in your homework problem proofs, take a minute at the end of each problem to reflect on where triangle inequality was motivated and how it moved you closer to your goal. This will help you significantly once you hit chapter 3.
- **Approach proofs calmly and strategically**. A lot of times, proofs can be very long, and problems can seem completely unapproachable. When you start dumping time into a problem throwing random tools at it without any motivation for why, there's a high chance you won't get anywhere. But if you spend the time to ground yourself, focus on what you have to prove, break it down into steps, and find tools that have the ability to move you closer to your end goal, you will struggle (and hopefully succeed) much more productively.

Good luck on the problem set!

8. Continuous Mappings

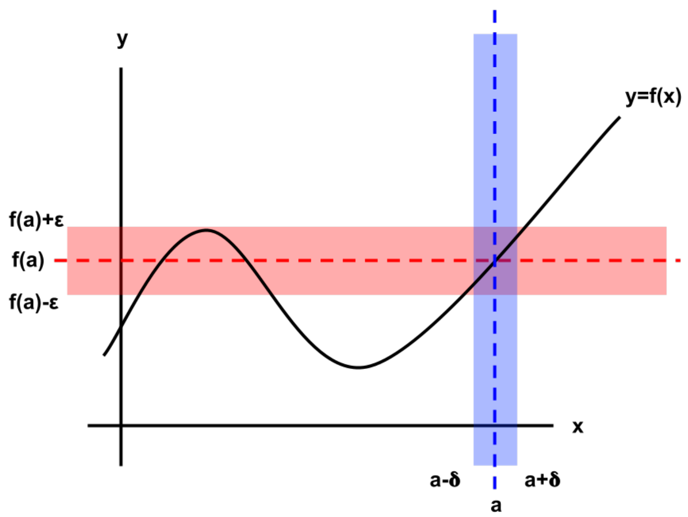
§8.1. Defining Continuity

We define continuity as follows:

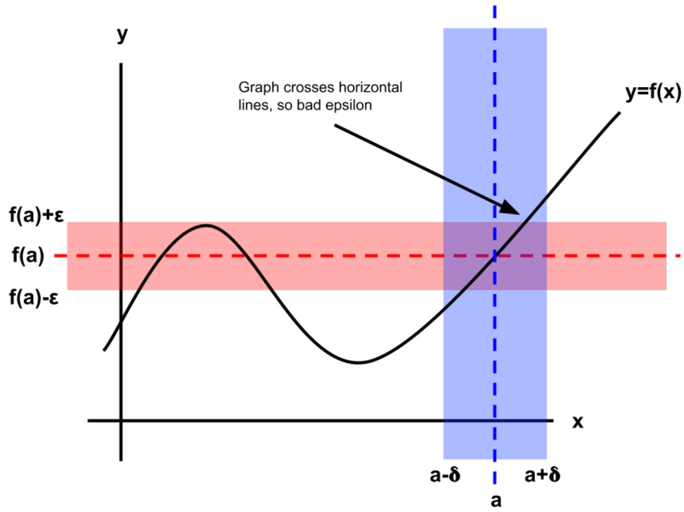
Definition 8.1.1

We have some function (or a well defined mapping) f from two metric spaces (S, ρ_S) and (T, ρ_T). We call a function **continuous** at a point $a \in S$ if for any real number $\varepsilon > 0$ there exists a real number $\delta > 0$ such that $\rho_s(x, a) < \delta$ implies $\rho_T(f(x), f(a)) < \varepsilon$.

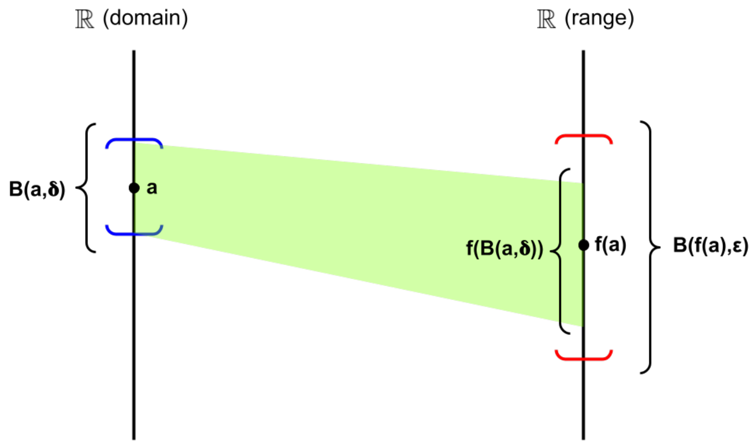
Lets see some pictures to understand visual ways to interpret this statement, looking at examples of functions from $\mathbb{R}$ to $\mathbb{R}$.



Above is one way to visualize what is going on the epsilon delta definition. Given any epsilon greater than zero, we get a little region in y (or the codomain of our function) colored in red, that we need to find a little region in x (of radius δ), colored in blue, such that $f(x)$ of all points in the blue region fit in the red region. Below is an example of a delta that fails, where the function $y = f(x)$ doesn't completely fit in the red region for all points in the blue region. You can see that if our function has an "infinitely steep part", then any δ value would fail.



Another equivalent way to define continuity is that a function f is continuous at a if for any real number $\varepsilon > 0$ there exists a real number $\delta > 0$ such that $f(B_S(a,\delta)) \subset B_T(f(a),\varepsilon)$ (make sure you see why this is true). There is a picture corresponding to this definition of continuity below (with a function from $\mathbb{R}$ to $\mathbb{R}$):



Warning: now that we've generalized the idea of continuity, the idea of continuity meaning "not picking your pencil up" no longer holds. For instance, the function

$$f(x) = \begin{cases} 0 & \text{if } x \in (0,1) \\ 1 & \text{otherwise} \end{cases}$$

may not be continuous as a function from $\mathbb{R}$ to $\mathbb{R}$, but it is as a function from $(-\infty,0) \cup (1,\infty)$ to $\mathbb{R}$.

§8.2. Epsilon Delta Proofs

Rigorously proving that a function is continuous uses the same idea for every proof: given some epsilon, you need to find a delta such that the conditions hold. Since you can't just manually pick a delta for every possible epsilon greater than zero (since there's uncountably many possibilities and that's also boring), we often want to find δ as a function of ε, such that for any corresponding ε we have a δ that fits the inequality required by continuity. Many of these proofs feel like you're working backwards: first, finding the "delta function", and then second, showing that this satisfies the requirements of continuity.

Example 8.2.1

Show that $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = x^2$ is continuous.

Alrighty, so what we need to do is make the quantity $|f(x) - f(a)| = |x^2 - a^2|$ small whenever $|x - a|$ is small. Why we want to do this is as we make our δ value smaller and smaller (which is the bound on the size of $|x - a|$), we want $|f(x) - f(a)|$ to get smaller and smaller, so that by choosing a small enough δ, we can make $|f(x) - f(a)|$ arbitrarily small and always get it smaller than some given ε. So we somehow want to relate $|x^2 - a^2|$ to $|x - a|$.

We have that $|x^2 - a^2| = |x - a||x + a|$, and we would like to bound these terms. We know $|x- < \delta$, and that $|x + a| \leq |x - a| + |2a| \leq \delta + |2a|$. This means that for a given $\varepsilon > 0$, $|x^2 - a^2| < \varepsilon$ whenever $\delta(\delta + |2a|) < \varepsilon$. This is kind of annoying, since we have a quadratic in δ and can't isolate it nicely. However, we don't care about equality here, so we can use cases and bounds to make our work easier. In the case that δ is less than 1, $\delta + 2|a| < 1 + 2|a|$, so when a given ε forces δ to be less than 1, we can just choose some $\delta \leq \frac{\varepsilon}{1+2|a|}$. In the case where the given ε doesn't force δ to be less than 1, we'll just pick $\delta = 1$ and we're guaranteed that this is a small enough δ. So given an ϵ, we just take δ to be $\min(1, \frac{\varepsilon}{1+2|a|})$, we're guaranteed that this δ will work.

Finally, we just need to show that our thought process actually works. This is how the proof you submit to a homework assignment would look: Suppose $\varepsilon > 0$. Then let $\delta = \min(1, \frac{\varepsilon}{1+2|a|})$. If $|x - a| < \delta$, then

$$\begin{aligned} |f(x) - f(a)| = |x^2 - a^2| &= |x - a||x + a| \\ &\leq |x - a|(|x - a| + |2a|) \\ &\leq \delta(\delta + |2a|) \qquad \leq \frac{\varepsilon}{1 + 2|a|}(1 + |2a|) = \varepsilon \end{aligned}$$

Take the time to understand what happened in this, particularly in the last two lines of the inequality rearranging. This might seem like it came out of nowhere, but it's just all the work we did earlier in guessing an appropriate δ, just written in reverse.

Exercise 8.2.2

Use a similar argument to show that $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = x^n$ is continuous, for any natural number n.

Showing that a function is discontinuous just requires finding a point where the inequality for continuity doesn't hold. For example, lets look at this function from earlier

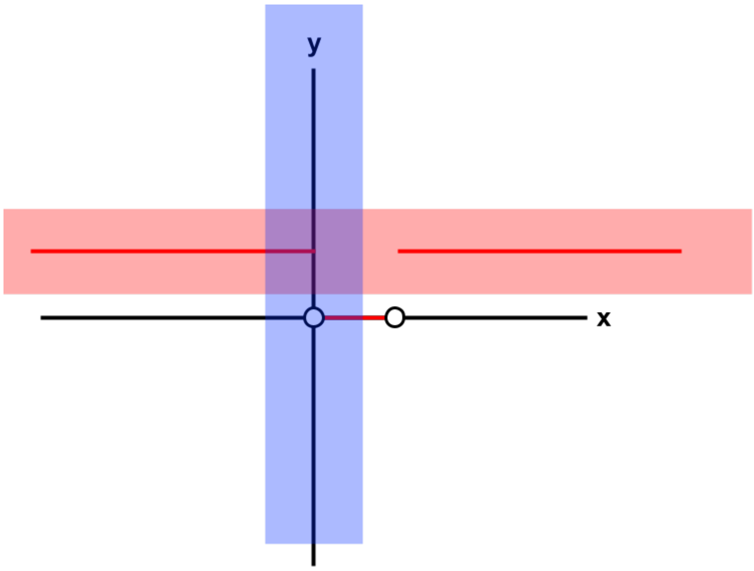
Example 8.2.3

Show rigorously that $f : \mathbb{R} \to \mathbb{R}$ such that

$$f(x) = \begin{cases} 0 & \text{if } x \in (0,1) \\ 1 & \text{otherwise} \end{cases}$$

is discontinuous.

Again, pictures always help.



If you look at line $y = 1$, notice that for any ϵ smaller than 1, lets say $\frac{1}{2}$, the ball $B(\varepsilon, \frac{1}{2})$ won't reach values of the function evaluated in the region $(0,1)$. So to rigorously say this function is discontinuous, we just have to find a single point, choose a single ϵ such that the continuity argument doesn't hold for any δ. We investigate the point $x = 0$, and for simplicity, we choose $\varepsilon = \frac{1}{2}$), then for all $\delta > 0$, there exists a point $a \in (0,1)$ contained in $(-\delta, \delta)$, and $|f(a) - f(0)| = |0 - 1| > \frac{1}{2}$.

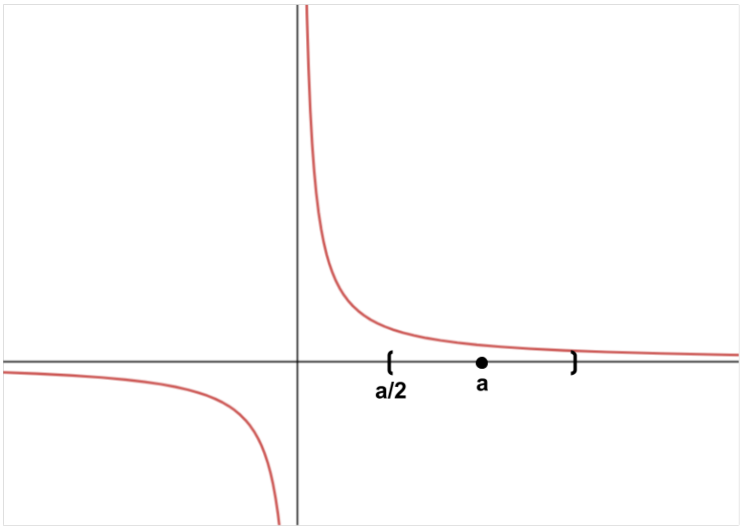
One challenging example that shows up in Gunning that he glosses over very quickly is the following, the last one we'll present in this section.

Example 8.2.4

Show rigorously that $f : \mathbb{R} \to \mathbb{R}$ such that $f(x) = \frac{1}{x}$ is continuous at any point $a \neq 0$.

Like before, we investigate $|f(x) - f(a)|$ (for $a \neq 0$) and try to show that it gets arbitrarily close to zero as $\delta > |x - a|$ gets arbitrarily close to zero. $|f(x) - f(a)| = |\frac{1}{x} - \frac{1}{a}| = |\frac{a-x}{xa}| = \frac{\delta}{xa}$. This means that for a given $\varepsilon > 0$, $|x^2 - a^2| < \varepsilon$ whenever $\frac{\delta}{xa} < \varepsilon$. However, the issue here is that δ cannot be dependent on x (where we are in our tiny δ interval we choose, it can only be independent on the location we choose a to be. We need to figure out how to bound $|xa|$.

The tricky thing to see here is that we want to bound $|xa|$ above. Although $\delta < \varepsilon|xa|$ and bounding xa independently of x below would give us that δ is less than something else independent of x, this is bad because δ needs to be strictly less than $\varepsilon|xa|$, so bounding it by term greater than $\varepsilon|xa|$ is a faulty argument. Instead, we need to bound $|xa|$ below, so that we can say that something independent of x is less than $\frac{\delta}{xa}$, which is less than ε. Let's draw a picture to get some ideas.



Picking some point a, we notice that whatever delta we pick must be less than or equal to a, or else we encompass the discontinuous point 0. So this leads to an idea where you take δ to be the minimum of a and whatever bound we get by working with the $\delta < a$ case. However, our issue is that as x approaches zero, $|ax|$ approaches zero too, so the only lower bound we can put on $|ax|$ is zero, which is useless to us. Instead, we work with the case where $\delta < \frac{a}{2}$, allowing us to bound $|ax|$ below by $|a||\frac{a/2}{|} = \frac{a^2}{2}$. So whenever ϵ forces δ to be less than $\frac{a}{2}$, we pick a δ such that $\frac{2}{a^2}\delta < \varepsilon$, and when this is not forced, we just pick $\delta = \frac{a}{2}$.

Exercise 8.2.5

Finish the rest of this example on your own, writing the needed δ as a minimum of two values (that may be dependent on ε, and showing that this δ expression guarantees continuity.

§8.3. Multivariable Continuity

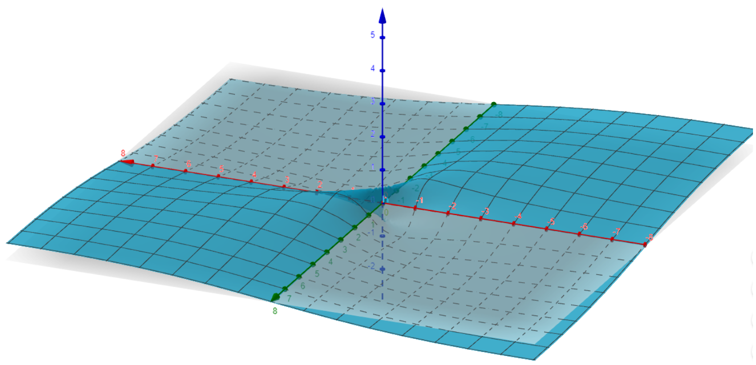
The definition of continuity for multivariable functions is exactly the same as for single variable functions, now that we've generalized the concept for continuity. As long as there is a defined metric in each space, there's no issue with checking if a multivariable function, say from $\mathbb{R}^3$ to $\mathbb{R}^2$, is continuous.

Let's say that we have a function $f : S \to \mathbb{R}^n$, where S is a metric space. We can break this function f into it's **coordinate functions** by investigating what f does to each of the n coordinates in $\mathbb{R}^n$, by viewing $f(x)$ as $(f_1(x), \ldots, f_n(x))$ (where each $f_i(x)$ is a function from S to $\mathbb{R}$. What is extremely important (and helpful to prove on your own) is that a function $f : S \to \mathbb{R}^n$ is continuous on S if and only if each coordinate function $f_i(x)$ is continuous on S.

Exercise 8.3.1

Prove this fact, using the l_∞ norm on $\mathbb{R}^n$ (we proved the l_∞ norm is equivalent to the standard l_2 norm in section 2.1, and equivalent norms/metrics preserve continuity).

This is great! We can turn any function with a multidimensional range into a bunch of functions with single dimensional range. The natural next question is to explore the continuity of a function $f : \mathbb{R}^n \to S$. Can we reduce the continuity of $f : \mathbb{R}^n \to S$ to the continuity of some kind of functions $f_i : \mathbb{R} \to S$? Unfortunately (or fortunately, depending how you look at it, since this would make multi really boring) we cannot. Example 3.7 in Gunning gives an example of a function from $\mathbb{R}^2$ to $\mathbb{R}$ **I GOTTA FINISH THiS LATER**



say something else here about quick mention of linearity?

§8.4. Sums and Products Continuous Mappings

The sum of two continuous functions being continuous is a straightforward exercise and we leave that to you to prove yourself.

Exercise 8.4.1

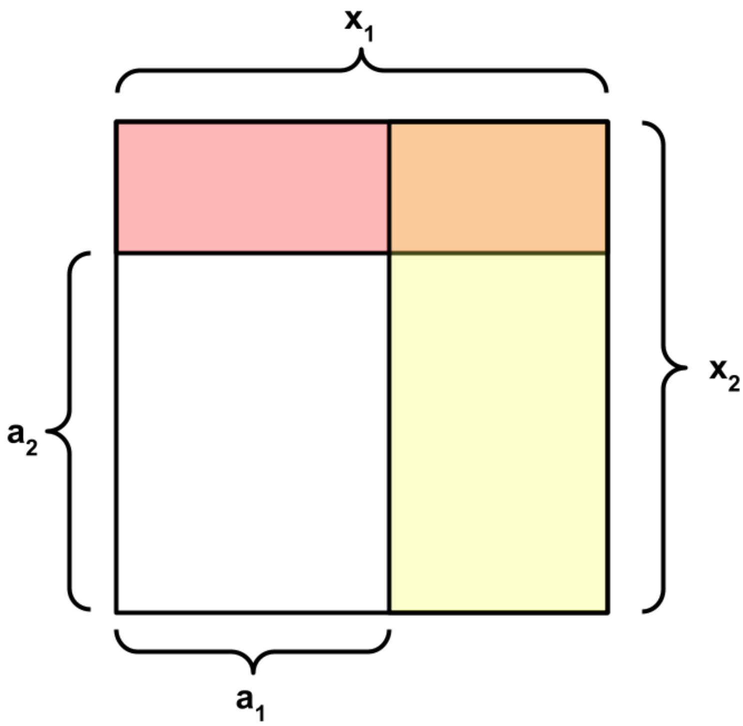
Prove that the sum of two continuous functions is continuous.

The proof that the product of two continuous functions is continuous isn't motivated particularly well in Gunning, so we'll show it here.

Theorem 8.4.2

The function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(x_1, x_2) = x_1 x_2$ is continuous (which is equivalent to showing the product of two continuous functions is continuous).

Proof. The idea, like always, is to relate $|f(\mathbf{x}) - f(\mathbf{a})| = |x_1x_2 - a_1a_2|$ to $|\mathbf{x} - \mathbf{a}|$. The expression $x_1x_2 - a_1a_2$ motivates us to draw a picture in order to find a neat way to play with and bound these products:



Note that we can represent the colored area (which is exactly $x_1x_2 - a_1a_2$) as $(x_2 - a_2)(x_1 - a_1) + a_1(x_2 - a_2) + a_2(x_1 - a_1)$ (you can foil this out if you'd like to make sure these two expressions are equal). Using the l_∞ norm on $\mathbb{R}^2$ tells us that whatever δ we pick will give us a little region $||\mathbf{x} - \mathbf{a}||_\infty$, where $\delta > |x_1 - a_1|$ and $|x_2 - a_2|$. So we have that

$$\delta^2 + a_1\delta + a_2\delta < |(x_2 - a_2)(x_1 - a_1) + a_1(x_2 - a_2) + a_2(x_1 - a_1)|$$

To get rid of the square on δ, we consider the case where $\delta < 1$, and then bound δ^2 above by δ. We encourage you to finish the rest of the proof.

Exercise 8.4.3

Complete the proof of this theorem (you should show that $\delta = \min(1, \frac{\varepsilon}{1+a_1+a_2})$ works). Also check out Gunning, who uses a slightly different δ.

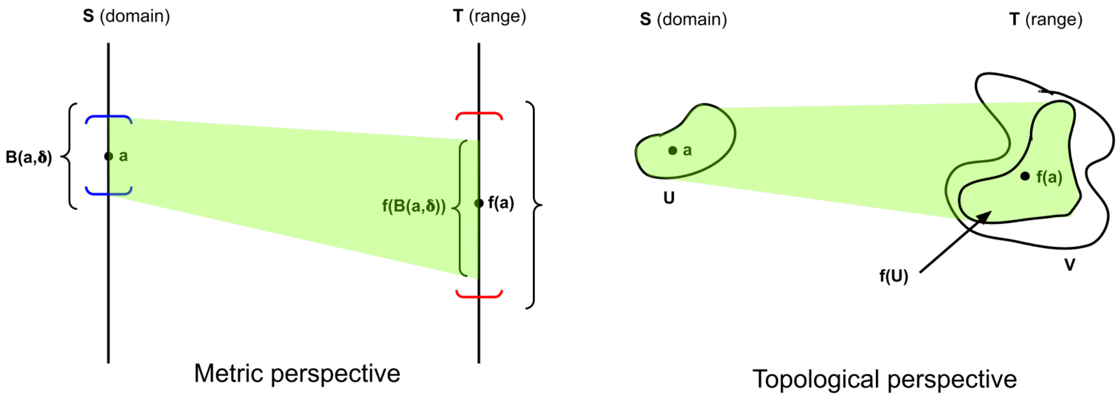
□

§8.5. The Return of Topologies

We've defined continuity from metric spaces to metric spaces, but we spent two whole sections focusing on point-set topology, so it would be kind of lame to not use any of it. So we are motivated to create a definition of continuity for maps between topological spaces, and try to relate it to our current definition of continuity between maps of metric spaces. Since all metric spaces are topological spaces, if we can show equivalence between continuity in a metric sense and a topological sense, we can exploit a huge amount of power from topology to prove things about analysis. Here's how we'll go about it. Let's go back to this view of continuity from earlier: a function $f : S \to T$ of metric spaces is continuous at a if for any real number $\varepsilon > 0$ there exists a real number $\delta > 0$ such that $f(B_S(a, \delta)) \subset B_T(f(a), \varepsilon)$. This definition is basically saying "for all neighborhoods

N_T (in the metric sense) of $f(a)$ in T there exists some neighborhood N_S of a in S such that $f(N_S) \subset N_T$. This notion is very easily extendable to topologies. Replacing the word "neighborhood" with "open set" allows us to treat S and T as topological spaces rather than metric spaces. So what we've done is define a notion of continuity between topological spaces that means the exact same thing as continuity between metric spaces when we're working with metric spaces: for a mapping $f : S \to T$ to be continuous at a point a means

$\forall$ open sets V containing $f(a)$ in T, $\exists$ an open set U containing a in S such that $f(U) \subset V$



Here's a quick exercise to familiarize yourself with this definition:

Exercise 8.5.1

Let S and T be metric spaces, and define a function $f : S \to T$. Show that the topological definition of continuity is equivalent to the following definition:

For all open sets V containing $f(a)$ in T, a is in the interior or $f^{-1}(V)$

Surprisingly, this topological continuity definition is equivalent to the following definition:

Theorem 8.5.2

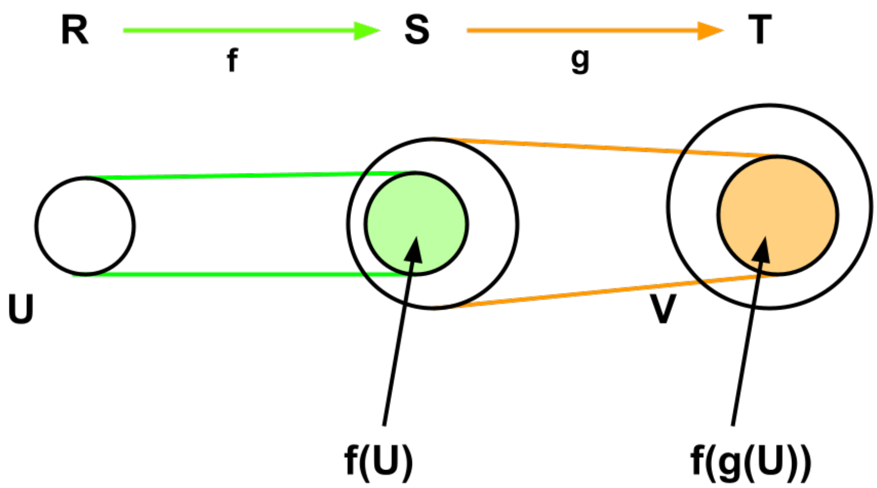
A mapping $f : S \to T$ is continuous if and only if the inverse image of any open set in T is open in S, if and only if the inverse image of any closed set in T is open in S.

Proving this is not hard, and again, is a great way to practice your topology skills in an analysis context.

Exercise 8.5.3

Prove the previous theorem, using the previous exercise as your definition for continuity (it helps a lot). Try to relate the interior of a set to showing whether or not a set is open.

And with the power of topologies, we can show ridiculously easily that the composition of continuous functions of topological function are continuous. Proof by picture:



Exercise 8.5.4
Explain this picture and why this suffices for the proof.

§8.6. The [insert name] Value Theorems

1) Let $\{U_\alpha\}$ be an open cover for $f(K)$
2) $\{f^{-1}(U_\alpha\}_\alpha$ is an open cover for K
3) There is a finite subcover $\{f^{-1}(U_{\alpha_i})\}_{i=1}^m$ for K
4) $\{U_{\alpha_i}\}_{i=1}^m$ is a finite subcover for $f(K)$

§8.7. Limits

§8.8. Uniform Continuity

§8.9. Uniform Convergence

§8.10. Wacky Examples

§8.11. Key Takeaways